طراحی سیستم هوشمند جهت پیش‌بینی بیماری‌های ژنتیکی کروموزومی با استفاده از تکنیک‌های داده‌کاوی

فریبا صلاحی[1*]، نسترن فرج پور[2]

تاریخ دریافت: ۱۴۰۰/۲/۱۱ تاریخ پذیرش: ۱۴۰۰/۱۰/۲۶

چکیده:

زمینه و هدف: امروزه شاهد پیشرفت‌های شگرف در زمینه داده‌کاوی داده‌های پزشکی هستیم. داده‌هایی که با تجزیه‌وتحلیل و کشف روابط بین آن‌ها می‌توان به الگوریتم‌هایی رسید که در پیشگیری و یا درمان بسیاری از بیماری‌ها به ما کمک می‌کنند. در این بین بیماری‌های ژنتیکی بخش اعظمی از توجه دنیای پزشکی را به خود اختصاص داده‌اند زیرا تولد کودکان مبتلا به ناهنجاری‌های ژنتیکی بار مالی، روانی و عاطفی زیادی به جامعه تحمیل می‌کند؛ بنابراین هدف این مطالعه ارائه الگوریتمی به عنوان یک آزمایش غربالگری ثانویه قبل از انجام تست‌های سلولی و مولکولی است.

مواد و روش‌ها: در این تحقیق ۱۰۰۰ پرونده مربوط به زنان بارداری که بعد از آزمایشات غربالگری در گروه ریسک متوسط یا بالا قرار داشتند مورد مطالعه قرار گرفت. اطلاعات بالینی آن‌ها ذخیره گردید، پاک‌سازی و حذف داده‌های و یکپارچه‌سازی رکوردها صورت گرفت سپس با استفاده از نرم‌افزار spss modeler، داده‌کاوی و کشف روابط بین داده‌ها صورت گرفت و در انتها الگوریتم مناسب برای شناسایی بیماری‌های ژنتیکی مشخص گردید.

نتایج: با اعمال پنج الگوریتم، شبکه‌های عصبی، ماشین بردار پشتیبانی، درخت تصمیم دودویی، درخت تصمیم چندتایی و رگرسیون لجستیک بر روی داده‌ها، مشخص گردید الگوریتم شبکه‌های عصبی با دقت ۹۷.۵۲۲٪ بیشترین میزان موفقیت را در تشخیص بیماری‌های ژنتیکی – کروموزومی قبل از تولد را دارا می‌باشد.

نتیجه‌گیری: استفاده از الگوریتم شبکه‌های عصبی به عنوان یک آزمایش غربالگری سبب می‌گردد که تعداد افراد کمتری کاندید انجام تست‌های پرهزینه و خطرناک سلولی و مولکولی قرار گیرند و می‌تواند به عنوان ابزاری در راستای کمک به کشف بیماری‌ها در دنیای پزشکی مورد استفاده قرار گیرد.

کلمات کلیدی: الگوریتم‌های کلاس‌بندی، شبکه‌های عصبی، ناهنجاری کروموزومی، ماشین بردار پشتیبان

[1] استادیار گروه مدیریت صنعتی، دانشکده مدیریت، دانشگاه آزاد اسلامی واحد الکترونیکی، تهران، ایران. (* نویسنده مسئول)
ایمیل: salahi_en@yahoo.com
[2] کارشناسی ارشد مدیریت سیستم های اطلاعاتی پیشرفته، دانشکده مدیریت، دانشگاه آزاد اسلامی واحد الکترونیکی، تهران، ایران

مقدمه:

از دیرباز جوامع بشری با بیماری‌های ژنتیکی لاعلاج مانند سندروم داون روبه رو بوده‌اند و راه‌حلی جز مدارا با این افراد وجود نداشت چون این افراد از بدو تولد با مشکلات عدیده‌ای همچون بیماری‌های قلبی عروقی، نقص سیستم ایمنی، تنفسی روبه‌رو هستند و از طرفی با توجه به پایین بودن سطح هوش (بین ۵۰-۷۰)، هیچ‌گاه توانایی مستقل زندگی کردن را ندارند و تنها بار مالی و عاطفی بر بدنه خانواده و جامعه وارد می‌کنند (۱). لذا پزشکان علم ژنتیک بر آن شدند تا با تشخیص زودهنگام این بیماری‌ها در دوران جنینی و ختم این‌گونه بارداری‌ها از تولد این‌گونه نوزادان جلوگیری کنند (۲).

سه روش جهت تشخیص زودهنگام قبل از تولد وجود دارد. ۱- نمونه‌برداری‌های آمنیوسنتز[1] با دقت ۹۹.۴٪ استخراج مقداری از مایع درون کیسه آب جنین به عنوان نمونه حاوی دی‌ان‌ای[2] (DNA) جنین، ۲- نمونه‌برداری از پرزهای جفتی[3] (CVS) با دقت ۹۸٪ (برداشت سلول از سطح جفت) و ۳- آزمایش خون غیرتهاجمی پیش از تولد[4] (NIPT) با دقت ۹۹٪ (اندازه‌گیری DNA آزاد جنین در خون مادر که از سلول‌های جفت آزاد می‌شوند). با استفاده از این سه روش به دی ان ای جنین دسترسی پیدا کرده و تشخیص می‌دهند جنین دچار ناهنجاری کروموزومی می‌باشد یا خیر (۳).

بنابراین پزشکان با توجه به برخی پارامترهای خونی مرتبط با بارداری آزمایشات غربالگری را طراحی کردند که بر اساس فرمولی بر پایه این پارامترها ریسک احتمال ابتلا را محاسبه می‌کند که بر اساس این ریسک زنان باردار در سه گروه ریسک پایین، ریسک متوسط و ریسک بالا تقسیم‌بندی می‌کنند (۴).

از بین بیماری‌های ژنتیکی– کروموزومی سندروم داون، پاتو، ادوارد، فریاد گربه، نقص لوله عصبی و هیدورسفالی به عنوان خطرناک‌ترین این بیماری‌ها شناخته می‌شود که معلولیت‌های شدید ذهنی و جسمی ایجاد می‌کند که پزشکان سعی در شناسایی زودهنگام آن‌ها و ختم بارداری این موارد می‌باشند (۵، ۶). حدود ۱۵-۱۰ درصد حاملگی‌های قطعی بالینی به سقط ختم می‌شود، این در حالی است که ۵۰٪ ختم حاملگی‌های زودرس به علت ناهنجاری‌های کروموزومی است (۷). شایع‌ترین ناهنجاری تشخیص داده‌شده تریزومی‌ها[5] می‌باشد (۶۱/۲ درصد) و در پی آن به ترتیب تریپلوئیدی[6] (۱۲/۴ درصد)، مونوزومی ایکس[7] (۱۰/۵ درصد) و ناهنجاری‌های ساختمانی کروموزوم‌ها (۴/۷ درصد) است (۸). در حال حاضر اصلی‌ترین علت انجام آزمایشات تهاجمی دوران بارداری (آمنیوسنتز یا نمونه‌برداری پرزهای جفتی) برای تشخیص ناهنجاری‌های کروموزومی است. هرچند بخاطر وجود ریسک سقط‌جنین، این آزمایشات فقط در حاملگی‌هایی انجام می‌شود که ریسک چنین آنوپلوئیدی‌هایی بالا باشد (۹،۱۰).

اختلالات کروموزوم جنسی تأثیر کمتری بر بقای جنین دارد و موجب ناباروری، نقص‌های مادرزادی از قبیل بیماری دریچه قلب در سندرم ترنر می‌شود (۱۱، ۱۲). اختلالات کروموزومی ممکن است در هر حاملگی رخ دهد، اما ریسک سندرم داون، ادوارد و پاتو با افزایش سن مادر بالا می‌رود (۱۳) لیست اختلالات کروموزومی شایع در جدول (۱) ارائه شده است.

جدول ۱: لیست اختلالات کروموزمی و ویژگی‌های بالینی آن‌ها

گروه	اختلال	تأثیر سن مادر	درصد موارد مرکب	سایر ویژگی‌های بالینی	منبع
اختلالات اتوزومال	سندرم داون (T21)	بله	۵۰-۴۰	مشکلات یادگیری، افت ایمنی، سالمندی زودرس	Hecht (1996), Morris et al. (2014)
	سندرم ادواردز (T18)	بله	۱۰۰	اختلال شدید رشد و تکامل، مرگ زودرس نوزادی	Savva et al. (2010)
	سندرم پاتو (T13)	بله			
اختلالات کروموزوم جنسی	سندرم ترنر (45,X)	ندارد	۳۰	بالغ نشدن، نازایی	Alberman and Creasy (1977),
	سندرم کلاین فیلتر	بعضی	-	هیپوگونادیسم، نازایی	Bojesen et al (2003)

[1] Amniocentesis
[2] Deoxyribonucleic acid
[3] Chorionic villus sampling
[4] Noninvasive prenatal testing
[5] Trisomy
[6] Triploid
[7] Monosomy X

در بسیاری از کشورها از جمله کشور ما حجم عظیمی از داده‌های مربوط به مراقبت و سلامت وجود دارد که هرگز مورد تحلیل قرار نگرفته‌اند لذا در بازار رقابتی امروز، سازمان‌هایی سریع‌تر به موفقیت دست پیدا می‌کنند که با استفاده از فناوری‌های جدید مانند داده‌کاوی بتوانند این انبوه اطلاعات را خلاصه، ذخیره و پردازش کنند و اطلاعات را استخراج و الگوهای حاکم بین داده‌ها را کشف کنند (۱۴-۱۶). در رابطه با داده‌کاوی در زمینه پزشکی تحقیقات بسیاری صورت گرفته است. غریبی (۱۳۹۵) تشخیص موارد ابتلا به سندروم داون را با استفاده از داده-کاوی ترکیبی موردبررسی قرار داد و با ترکیب مدل‌های درخت تصمیم و بیز ساده دقت ۹۶.۶۷٪ در تشخیص این سندروم را به دست آورد (۱۷). میرزایی و همکاران (۱۳۹۵) کاربرد داده‌کاوی در پیش‌بینی بقای پیوند کلیه و شناسایی متغیرهای تأثیرگذار در بقای کلیه پیوندی را بررسی کردند آن‌ها از طبقه‌بندی‌های شبکه عصبی، درخت تصمیم و ماشین‌بردار پشتیبان جهت پیش‌بینی بقای پیوند کلیه استفاده کردند و نتیجه گرفتند با استفاده از همجوشی اطلاعات، می‌توان صحت نتایج طبقه‌بندی‌ها را افزایش داد (۱۸).

مویرا و نامن[1] (۲۰۱۸) یک مدل داده‌کاوی پیوندی برای تشخیص بیماران مبتلا به شک بالینی جنون پیشنهاد کردند و هدف آن کمک به متخصصان در تشخیص بیماران مبتلا به شک بالینی جنون بوده است (۱۹). در سال ۲۰۱۸ سیمیک[2] و همکاران رویکرد خوشه‌بندی پیوندی را برای تشخیص بیماری‌های طبی بررسی کردند. این مطالعه روی ایجاد یک راهبرد جدید بر پایه مدل پیوندی برای ترکیب روش پارتیشن‌بندی فازی و الگوریتم خوشه‌بندی برآورد حداکثر احتمال جهت تشخیص بیماری‌ها تمرکز داشت (۲۰). پاسانیسی و پایانویک[3] (۲۰۱۸) رویکرد داده‌کاوی اطلاعات پیوندی را برای کشف دانش در مورد بیماری‌های قلبی عروقی مورد آزمایش قراردادند. در این مطالعه خوش بندی، قواعد همبستگی و شبکه‌های عصبی برای بررسی و شناخت ریسک فاکتورهای واقعه قلبی با هدف کاهش ریسک بیماری قلبی و عروقی ترکیب کردند (۲۱). ژو[4] و همکاران (۲۰۱۸) به بررسی الگوریتم فورست کلاس وزن های تصادفی برای پردازش داده‌های پزشکی دارای کلاس نامتعادل پرداختند (۲۲).

لذا در این تحقیق بر آن شدیم به کمک علم داده‌کاوی الگوریتمی پیدا کنیم تا با استفاده از فاکتورهای استخراج شده از آزمایشات غربالگری، سونوگرافی، سوابق بالینی، به دور از خطاهای رایج انسانی طیف بهتری از افرادی که آزمایشات غربالگری آن‌ها دارای ریسک بالا بوده را شناسایی کرده تا افراد کمتری متحمل هزینه‌های گزاف و یا خطرات احتمالی تست‌های تهاجمی و غیرتهاجمی شوند. برای داده‌کاوی در علوم پزشکی تکنیک‌ها، رویکردها، الگوها و روش‌های مختلفی پیشنهاد شده است با توجه به ماهیت طبقه‌بندی داده‌ها الگوریتم‌های کلاس‌بندی برای انجام تحقیق موردنظر مناسب می‌باشند.

مواد و روش‌ها:

این پژوهش، کاربردی و از نوع مدل‌سازی بوده است. پس از تأیید علمی طرح و ضرورت انجام آن در گروه تخصصی و کمیته اخلاق، ۱۰۰۰ پرونده مربوط به بیمارانی که آزمایشات غربالگری آن‌ها دارای ریسک بالا بوده و برای انجام تست آمنیوسنتز به آزمایشگاه ژنتیک نیلو – ژنوم مراجعه کرده بودند، از پایگاه داده مربوطه به صورت تصادفی انتخاب گردیده و به بررسی فاکتورها و متغیرهای تعیین‌کننده در بروز این بیماری‌ها پرداخته شده است. پیش از شروع کار از مراجعه‌کنندگان رضایت‌نامه کتبی مبنی بر استفاده از نتایج و تفاسیر آزمایشات آن‌ها در راستای انجام تحقیقات نظری دریافت شد. پس از لیست کردن متغیرها، به مرتب‌سازی و نرمال‌سازی نمونه‌ها پرداخته شد و در نهایت از الگوریتم‌های مختلف داده‌کاوی به پیش‌بینی نتیجه اقدام شد و در نهایت الگوریتم با بیشترین دقت انتخاب، تست و اعتبارسنجی گردید. با توجه به ماهیت پژوهش که توصیف مارکرهای آزمایشات غربالگری و پیدا کردن الگویی از روابط بین آن‌هاست روش تحقیق توصیفی می‌باشد و از آنجا که همه پارامترها عددی بوده و پژوهش ارزیابانه می‌باشد، ماهیت پژوهش از نوع کمی می‌باشد.

فرآیند انجام پژوهش شامل گام‌های زیر می‌باشد.

۱- جمع‌آوری داده‌ها

۲- پیش‌پردازش

۳- پردازش

۴- مدل‌سازی

۱- **جمع‌آوری داده‌ها:** بعد از مشخص شدن متغیرهای پژوهش مطابق جدول ۲ داده‌ها جمع‌آوری و در پایگاه داده ذخیره گردیده‌اند جهت تعیین متغیرهای تحقیق از

[1] Moreira and Namen
[2] Simic
[3] Pasanisi and paiano
[4] Zhu

مقاله غریبی با عنوان تشخیص بیماری‌های سندوم دان با استفاده از داده‌کاوی و نظرات پزشکان و مشاوران ژنتیک استفاده شد.

۲- **مرحله پیش پردازش:** مهم‌ترین فعالیت‌های که در بخش پیش‌پردازش داده‌ها انجام می‌شود عبارت است از:

۲-۱ پاک‌سازی داده‌ها: مهم‌ترین فعالیت‌های این بخش عبارت است تخمین مقادیر ناموجود[1] در پایگاه داده‌ها، از بین بردن اختلال[2] در داده‌ها، حذف کردن داده‌های پرت و نامربوط، از بین بردن ناسازگاری در داده‌ها

۲-۲ یکپارچه‌سازی داده‌ها: در بسیاری از موارد ممکن است داده‌ها در فایل‌ها و منابع مختلف نگهداری شوند و در این صورت نیاز است تا داده‌ها پیش از اجرای تکنیک‌های داده‌کاوی با یکدیگر یکپارچه شوند.

۲-۳ کاهش داده‌ها: ممکن است همیشه، همه داده‌ها موردنیاز نباشند و تنها بخشی از داده‌ها که موردنیاز است باید مورد پردازش قرار بگیرد

۲-۴ تبدیل داده‌ها: فعالیت‌های مانند نرمال‌سازی داده‌ها و گسسته سازی داده‌ها در این حوزه جای می‌گیرند

۳- **پردازش داده‌ها**: پس از انجام عمل پیش‌پردازش بر روی متغیرها و حذف متغیرهای ناکارآمد، اکنون داده‌ها آماده استخراج الگو و مدل مناسب به منظور کشف دانش می‌باشند. با استفاده از روش‌های دسته‌بندی صفات تشخیص داده می‌شوند، داده‌ها دسته‌بندی می‌شوند و در نهایت با استفاده از الگوریتم‌های مختلف و مقداردهی آن‌ها با متغیرهای توضیحی (ورودی) مدل با بیشترین درصد تشخیص شناسایی می‌گردد

۴- **مدل‌سازی**: در نهایت با spss modeler مدل‌سازی داده‌ها انجام می‌گیرد.

جدول ۲: متغیرهای جمع‌آوری‌شده در پژوهش

نام متغیر	علامت اختصاری	نحوه استخراج
سن مادر	Age	پرسشنامه
دیابت	Diabetic	پرسشنامه
سیگار	Smoker	پرسشنامه
سابقه اختلال خانوادگی	Family history	پرسشنامه
سابقه اختلال در فرزندان قبلی	Partners history	پرسشنامه
هورمون بتا اچ سی جی آزاد	FBHCG- MOM FBHCG	آزمایشات غربالگری
پروتوئین A پلاسما مربوط به بارداری	PaPP-A - MOM PaPP-A	آزمایشات غربالگری
ریسک ابتلا	EDD RISK, PATU RISK, DS RISK	آزمایشات غربالگری
چین پشت گردنی	NT	سونوگرافی
سن بارداری	P-age	سونوگرافی
نتیجه	Result	آزمایش سلولی و مولکولی

[1] Missing value
[2] Noise

یافته‌ها

بنابر هدف کلی که ارائه الگوریتمی به عنوان یک آزمایش غربالگری قبل از انجام تست‌های سلولی و مولکولی تعداد افراد کمتری کاندید انجام تست های پرهزینه و خطرناک سلولی و مولکولی قرار گیرند و ارائه برترین مدل داده‌کاوی در تشخیص سندروم‌های داون، ادوارد و پاتو می‌باشد.

در این مطالعه داده‌های مشخصات فردی و بالینی و نتایج آزمایش ۱۰۰۰ مادر حامله مشکوک به تریزومی از پایگاه داده آزمایشگاه نیلو-ژنوم استخراج شد و به عنوان متغیرهای پژوهش مورد استفاده قرار گرفت. متغیرهای دموگرافیک شامل سن (سال)، سن حاملگی (هفته)، سابقه بیماری کروموزومی در خانواده زوجین، سابقه بیماری کروموزومی در جنین‌های قبلی زوجین و متغیرهای آزمایشگاهی شامل میزان پروتئین PAPPA، هورمون FBHCG، شفافیت پشت گردنی جنین، استخراج و جمع‌آوری گردید. همچنین یک متغیر Result به عنوان نتیجه خروجی در نظر گرفته شد. این متغیر به عنوان متغیر پاسخ (پیش‌بینی شونده) و باقی متغیرها به عنوان متغیر توضیحی (پیش‌بینی کننده) بوده‌اند. در نهایت با نرم‌افزار IBM SPSS Modeler مدل‌سازی داده‌ها انجام شد. مراحل جمع‌آوری اطلاعات تا تست مدل ایجادشده به صورت زیر بوده است:

جمع‌آوری داده‌ها

جدول ۳ توزیع فراوانی داده‌های مربوط به ریسک سندروم های سه‌گانه را نشان می‌دهد.

جدول ۳: توزیع فراوانی ریسک سندرم‌های سه‌گانه تریزومی بین واحدهای پژوهش

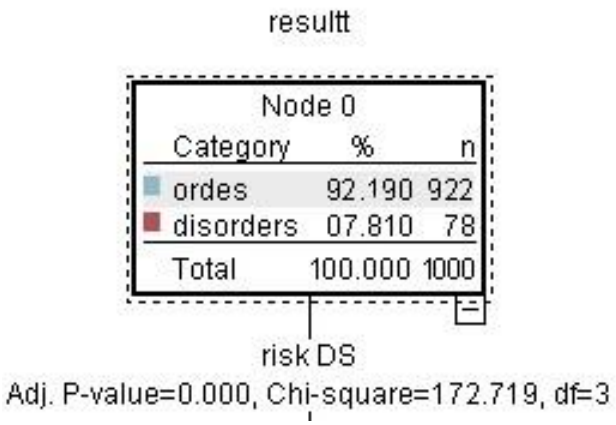
	سندروم داون		سندروم ادوارد		سندروم پاتو	
	فراوانی	درصد	فراوانی	درصد	فراوانی	درصد
ریسک بالا	۷۸	۱۵/۶	۱۳	۲/۶	۱۳	۲/۶
ریسک بینابینی	۳۴۸	۶۹/۶	۴۳	۸/۶	۵۵	۱۱
ریسک پایین	۷۴	۱۴/۸	۴۴۴	۸۸/۸	۴۳۲	۸۶/۴
جمع	۵۰۰	۱۰۰	۵۰۰	۱۰۰	۵۰۰	۱۰۰

پیش‌پردازش:

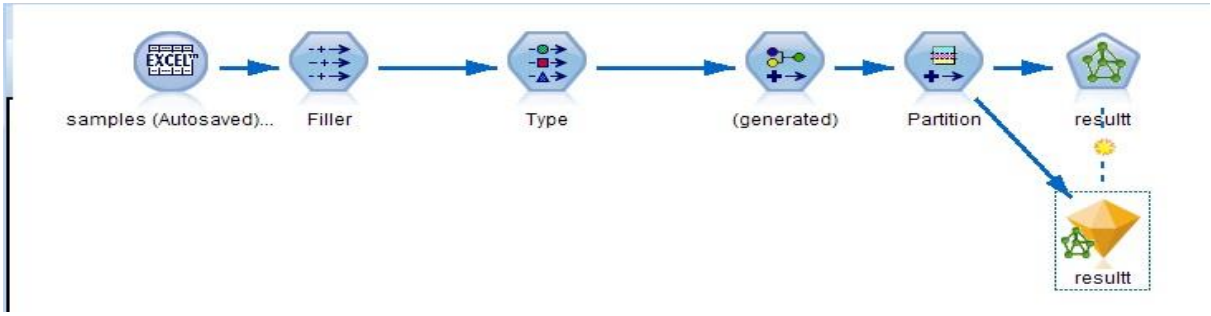
در هنگام گردآوری پایگاه داده‌ها فیلدهایی که داده‌های مفقودی[1] داشتند کاملاً حذف گردیدند پس از حذف رکوردهای حاوی داده‌های مفقودی، داده‌های پرونده‌های جدید جایگزین گردید همچنین متغیرهای بالینی مانند دیابت و سیگاری بودن به دلیل نویز[2] بالا و عدم اطمینان به‌راستی گفتار فرد آزمایش دهنده از لیست متغیرها حذف گردید سپس برچسب[3] و ارزش رتبه[4] هر یک از متغیرها تعریف شد. هم‌زمان شاخص‌های پراکندگی و مرکزی و توصیف داده‌ها انجام شد.

مدل‌سازی:

در این مطالعه داده‌ها با نسبت ۷۰ به ۳۰ درصد برای داده های تعلیمی و داده‌های تست تقسیم شد سپس الگوریتم‌های دسته‌بندی شامل الگوریتم C5، SVM، CHAID، Logestic Regression(LR)، Neural Network(NN) بر روی داده‌ها اجرا شد. به عنوان نمونه استریم اجراشده شبکه‌های عصبی در شکل ۱ قابل مشاهده می‌باشد.

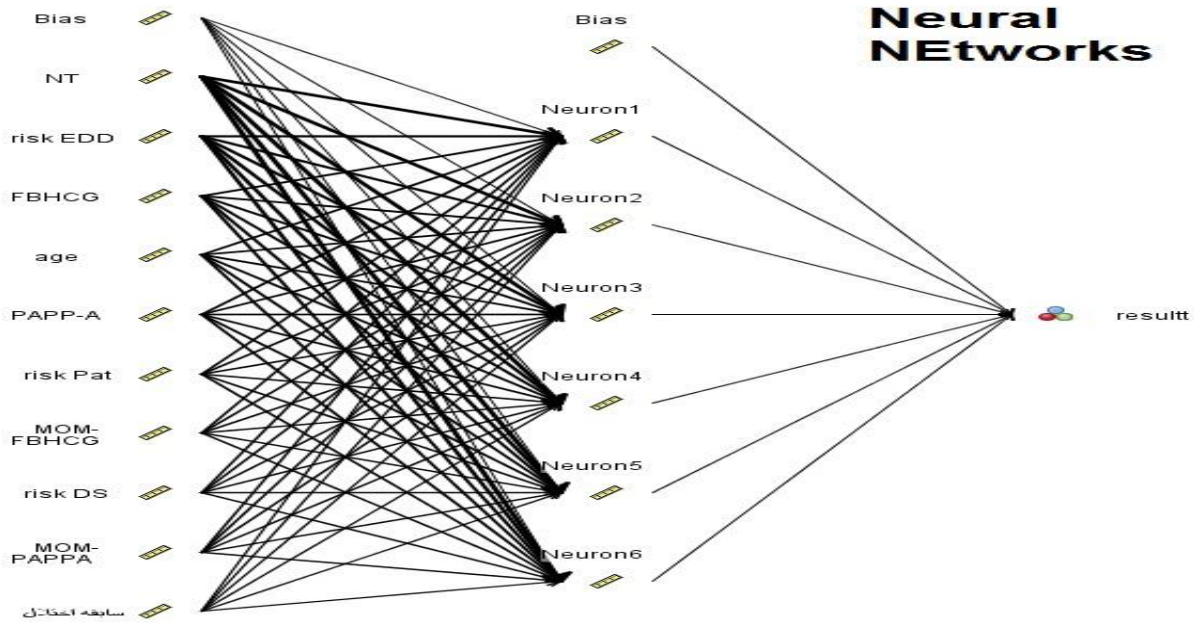


[1] Missing value
[2] Noise
[3] Lable
[4] Valu



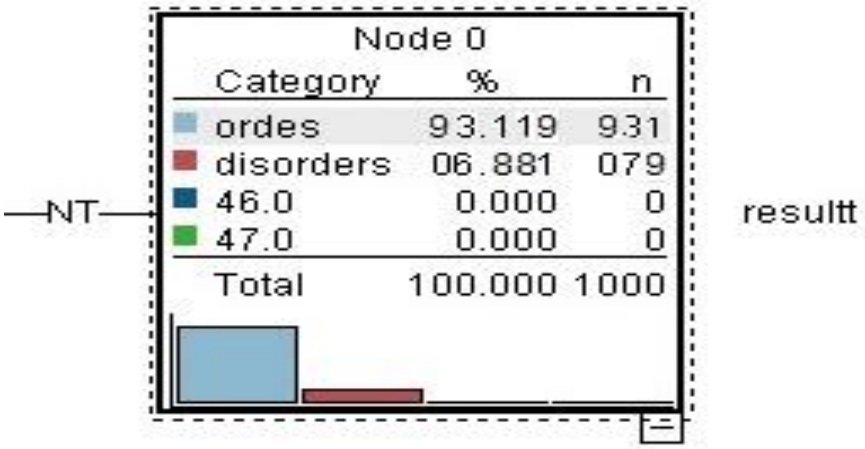
شکل ۱: استریم اجراشده شبکه‌های عصبی

در ادامه نتایج اجرای هر یک از الگوریتم‌ها به تفکیک استریم بیان شده است. شکل ۲ اجرای مدل شبکه عصبی با ۹۷/۵۲۲ درصد صحت تجزیه‌وتحلیل نشان می‌دهد که در ان سطح β_{HCG} آزاد سرم بیشترین قدرت پیش‌بینی ریسک ابتلای به تریزومی را دارد.



شکل ۲: تشکیل جدول نتایج کلاس‌بندی شبکه‌های عصبی

جدول ۴: نتایج کلاس‌بندی C5.0



جدول ۴ نشان می‌دهد در الگوریتم دسته‌بندی C5.0 متغیر قطر چین پشت گردن جنین (NT) با صحت ۹۳/۱۱۹ درصد بالاترین قدرت پیش‌بینی ابتلای به تریزومی را دارد. مهم‌ترین متغیر پیشگو در این الگوریتم چین پشت گردن بود.

جدول ۵ نشان می‌دهد که در الگوریتم دسته‌بندی CHAID متغیر میزان ریسک سندروم داون RISK DS با صحت ۹۲/۱۹۰ درصد بالاترین قدرت پیش‌بینی ابتلای به تریزومی را دارد. مهم‌ترین متغیر پیشگو در این الگوریتم مقدار ریسک سندروم داون بود.

جدول ۵: نتایج کلاس‌بندی CHAID

Results for output field result

Comparing $C-result with result

'Partition'	1_Training		2_Testing	
Correct	0	0%	0	0%
Wrong	687	100%	293	100%
Total	687		293	

Coincidence Matrix for $C-result (rows show actuals)

'Partition' = 1_Training	disorders	ordes
46	12	595
47	21	59

'Partition' = 2_Testing	disorders	ordes
46	11	246
47	11	25

جدول ۶: نتایج کلاس‌بندی SVM

Results for output field resultt

Comparing $S-resultt with resultt

'Partition'	1_Training		2_Testing	
Correct	653	93.29%	277	96.458%
Wrong	47	6.71%	23	7.67%
Total	700		300	

Results for output field result

Comparing $C-result with result

'Partition'	1_Training		2_Testing	
Correct	0	0%	0	0%
Wrong	687	100%	293	100%
Total	687		293	

Coincidence Matrix for $C-result (rows show actuals)

'Partition' = 1_Training	disorders	ordes
46	11	596
47	34	46

'Partition' = 2_Testing	disorders	ordes
46	14	243
47	14	22

جدول ۷: نتایج کلاس‌بندی LR

Results for output field resultt

Comparing $L-resultt with resultt

'Partition'	1_Training		2_Testing	
Correct	28	89.71%	267	90.196%
Wrong	72	10.29%	33	11%
Total	700		300	

جداول ۶ و ۷ میزان پیش‌بینی الگوریتم ماشین بردار پشتیبان و رگرسیون لجستیک نشان می‌دهند.

پنج الگوریتم کلاس‌بندی NN,SVM,C5,Chaid,LR که روی پایگاه داده موردنظر پیاده‌سازی گردیدند ابتدا نتایج در قالب دو کلاس تفکیک گردیده، کلاس بیمار (تشخیص صحیح تعداد بیماران از کل بیماران) کلاس سالم (تشخیص صحیح سالم‌ها از کل سالم‌ها) و در نهایت دقت کلی ارائه گردید. جدول ۸ مقایسه نتایج عملکرد الگوریتم‌ها را نشان می‌دهد.

جدول ۸ نتایج پیاده‌سازی الگوریتم‌های متفاوت

الگوریتم	دقت در کلاس بیمار	دقت در کلاس سالم	دقت کل
NN	95.228	98.915	97.522
SVM	95.189	97.389	96.458
LR	89.129	91.879	90.196
C5.0	91.968	94.225	93.119
CHAID	90.489	93.896	92.190

بحث:

ناهنجاری کروموزومی ناشی از تغییرات در ساختار و یا تعداد کروموزم‌ها می‌باشند. تریزومی‌ها به دلیل وجود یک کپی اضافی از کروموزوم به وجود می‌آیند بدین معنا که افراد مبتلا به این‌گونه ناهنجاری‌ها به جای ۴۶ کروموزوم، دارای ۴۷ کروموزوم می‌باشند که این ناهنجاری منجر به عقب‌افتادگی شدید ذهنی و جسمی کودکان می‌شوند پزشکان برای جلوگیری از به دنیا آمدن کودکان مبتلا به این‌گونه ناهنجاری‌ها، تست‌های پیشرفته سلولی مولکولی انجام می‌دهند که هم هزینه بسیار بالایی را به افراد تحمیل می‌نماید و هم جزو گروه تست‌های تهاجمی قرار دارند. الگوریتم‌های شبکه عصبی (که در این مطالعه بالاترین میزان دقت در تشخیص ابتلای جنین به ناهنجاری‌های ژنتیکی را دارا می‌باشد) می‌توانند به عنوان یک آزمایش غربالگری قبل از انجام تست‌های سلولی و مولکولی انجام گیرد تا با غربالگری طیف گسترده از نمونه‌های ارجاعی تعداد افراد کمتری را کاندید انجام تست‌های پرهزینه و خطرناک سلولی و مولکولی قرار دهد.

در پژوهش‌های پیشین داده‌کاوی به کمک بسیاری از شاخه‌های پزشکی آمده و متخصصان را قادر ساخته الگوریتم تشخیص‌های دقیقی را بسازند اما در حوزه تشخیص بیماری‌های ژنتیکی یا تمرکز بر تشخیص صرفاً سندروم داون بوده و باقی سندروم‌ها موردتوجه قرار نگرفته‌اند و یا تنها از یک الگوریتم از پیش تعیین‌شده به بررسی موضوع پرداخته‌اند اما در مطالعه حاضر هر ۳ سندروم خطرناک با استفاده از ۵ الگوریتم مختلف کلاس‌بندی شده و نتایج آن‌ها با یکدیگر مقایسه شده تا دقیق‌ترین الگوریتم انتخاب گردد. یافته‌های مطالعه نشان داد ۱۵۶ نفر (۱۵/۶ درصد) از واحدهای پژوهش در ریسک بالای سندرم داون، ۲۶ نفر (۲/۶ درصد) در ریسک بالای ابتلا به سندرم ادوارد و ۲۶ نفر (۲/۶ درصد) در ریسک بالای ابتلا به سندرم پاتو می‌باشند. بیشترین ریسک بینابینی در گروه سندرم داون مشاهده می‌شود این در حالی است که کمترین ریسک ابتلا به اختلالات تریزومی در گروه سندرم ادوارد با فراوانی ۸۸/۸ درصد بوده است. در این مطالعه متغیرهای پیشگو از بانک اطلاعاتی آزمایشگاه نیلو استخراج

شده و متغیر پیامد (پیش‌بینی) با استفاده از الگوریتم‌های خوشه‌بندی و دسته‌بندی آن‌ها با استفاده از الگوریتم‌های فوق‌الذکر محاسبه گردید. نتیجه اجرای مدل حاکی از حساسیت و اختصاصیت بالای مدل در پیش‌بینی ریسک تریزومی‌های چهارگانه مورد مطالعه بود.

این مطالعه به آزمون پیش‌فرض‌هایی بر اساس پیش‌پردازش داده‌های استخراج شده از پایگاه داده آزمایشگاه نیلو پرداخته است؛ بنابراین مطالعه تک مرکزی بوده و جامعه مراجعه‌کننده می‌تواند از یکدستی خاص برخوردار باشد؛ که برای رفع آن می‌توان داده‌ها را از چند مرکز و از چند دوره زمانی استخراج کرد. همچنین در پاک‌سازی و آماده‌سازی داده‌ها از متغیر زمان استفاده نشده است و می‌تواند از محدودیت‌های مطالعه باشد.

نتیجه‌گیری:

الگوریتم شبکه‌های عصبی با دقت ٪۹۷.۵۲۲ بهترین نتیجه را در تشخیص بیمار و یا سالم بودن کیس ارجاعی نشان داد که با استفاده از این الگوریتم می‌توان سیستم خبره و تصمیم یاری را طراحی کرد که با استفاده از فاکتورهای ورودی تا دقت ٪۹۷.۵۲۲ به درستی ابتلای جنین به بیماری‌های ژنتیکی – کروموزومی را تشخیص دهد و با اضافه کردن کیس‌های جدید، بخش یادگیری ماشین به دانش خود اضافه کرده و دقت تشخیص را بالا می‌برد. با توجه به این نتایج پیشنهاد می‌شود

ازآنجایی‌که آزمایشات سلولی و مولکولی خود دارای درصد خطاست از طرفی خطاهای انسانی در اعلام نتایج نیز بی‌تأثیر نمی‌باشد پیشنهاد می‌شود در یک بازه زمانی خاص مانند ۱ یا ۲ سال تمامی افراد مراجعه‌کننده آزمایشات غربالگری و سونوگرافی را در یک مرکز واحد انجام دهند تا خطای ناشی از تفاوت دستگاه‌ها و اپراتورها به حداقل برسد پس از تست غربالگری افرادی که در رنج ریسک متوسط یا ریسک بالا قرار می‌گیرند شناسایی‌شده و تست‌های سلولی و مولکولی توسط همان مرکز با شرایط یکسان روی این افراد صورت گیرد و بعد از تطابق نتایج این تست‌ها با نتیجه نهایی زایمان و یا سقط، یک پایگاه داده جامع تشکیل داده تا در اختیار الگوریتم‌های کلاس‌بندی شبکه‌های عصبی قرار گیرد آنگاه این الگوریتم با استفاده از پایگاه داده‌ها که حاوی تعداد بسیار زیادی داده آموزشی می‌باشد تعلیم و تست انجام می‌دهد و با پیاده‌سازی این الگوریتم و مقایسه نتایج آن با نتایج نهایی تست‌های سلولی و مولکولی نتایج بسیار دقیق‌تری ارائه خواهد داد همچنین این مطالعه را می‌توان با فاکتورهای بیوشیمی غربالگری دوم مانند آلفا فیتو پروتوئین [1]، اچ سی جی [2] واستریول غیرمزدوج [3] و هورمون اینهیبین آ [4] نیز تکرار کرد تا بتوان تشخیص داد کدام عوامل بالاترین میزان دقت را در نتایج تست ارائه می‌دهند.

تشکر و قدردانی:

نویسندگان این پژوهش از تمامی مدیران و کارکنان آزمایشگاه نیلو – ژنوم که نهایت همکاری که با این پروژه تحقیقاتی انجام دادند و باعث تسهیل کار پژوهش شدند کمال تشکر و قدردانی را دارند.

[1] AFP
[2] HCG
[3] UE3
[4] Inhibin A

References

1- Gardner RM, Sutherland GR, Shaffer LG. Chromosome abnormalities and genetic counseling. OUP USA; 2011 Nov 11.

2- Skotko BG, Levine SP, Goldstein R. Having a son or daughter with Down syndrome: Perspectives from mothers and fathers. American Journal of Medical Genetics Part A. 2011 Oct; 155(10):2335-47.

3- Aryan Z, Bahadori A, Farhud D. Prenatal diagnostic tests of genetic disorders. Tehran University Medical Journal. 2019; 77(1):8-12.

4- Shiefa S, Amargandhi M, Bhupendra J, Moulali S, Kristine T. First trimester maternal serum screening using biochemical markers PAPP-A and free β-hCG for Down syndrome, patau syndrome and edward syndrome. Indian Journal of Clinical Biochemistry. 2013 Jan 1; 28(1):3-12.

5- Carey JC. Trisomy 18 and trisomy 13 syndromes. Cassidy and Allanson's Management of Genetic Syndromes. 2021 Feb 19:937-56.

6- Cereda A, Carey JC. The trisomy 18 syndrome. Orphanet journal of rare diseases. 2012 Dec; 7(1):1-4.

7- Jenderny J. Chromosome aberrations in a large series of spontaneous miscarriages in the German population and review of the literature. Molecular cytogenetics. 2014 Dec; 7(1):1-9.

8- Soler A, Morales C, Mademont-Soler I, Margarit E, Borrell A, Borobio V, Muñoz M, Sánchez A. Overview of chromosome abnormalities in first trimester miscarriages: a series of 1,011 consecutive chorionic villi sample karyotypes. Cytogenetic and genome research. 2017; 152(2):81-9.

9- Ashoor Al Mahri G, Nicolaides KH. Evolution in screening for Down syndrome. The Obstetrician & Gynaecologist. 2019 Jan; 21(1):51-7.

10- Gray KJ, Wilkins-Haug LE. Have we done our last amniocentesis? Updates on cell-free DNA for Down syndrome screening. Pediatric radiology. 2018 Apr; 48(4):461-70.

11- Tartaglia N, Howell S, Wilson R, Janusz J, Boada R, Martin S, Frazier JB, Pfeiffer M, Regan K, McSwegin S, Zeitler P. The eXtraordinarY Kids Clinic: an interdisciplinary model of care for children and adolescents with sex chromosome aneuploidy. Journal of Multidisciplinary Healthcare. 2015; 8:323.

12- Pinsker JE. Turner syndrome: updating the paradigm of clinical care. The Journal of Clinical Endocrinology & Metabolism. 2012 Jun 1; 97(6):E994-1003.

13- Moorthie S, Blencowe H, Darlison MW, Gibbons S, Lawn JE, Mastroiacovo P, Morris JK, Modell B. Chromosomal disorders: estimating baseline birth prevalence and pregnancy outcomes worldwide. Journal of community genetics. 2018 Oct; 9(4):377-86.

14- Moghaddassi H, Hoseini A, Asadi F, Jahanbakhsh M. Application of Data Mining. Health Information Management 2012; 9(2): 304

15- Wright A, Chen ES, Maloney FL. An automated technique for identifying associations between medications, laboratory results and problems. Journal of biomedical informatics. 2010 Dec 1; 43(6):891-901.

16- Canlas RD. Data mining in healthcare: Current applications and issues. School of Information Systems & Management, Carnegie Mellon University, Australia. 2009 Aug 5.

17- Gharibi J , Diagnosis of Down syndrome using a combined data mining model. New research in engineering sciences. Alame Rafiee,conference. 2016 June.

18- Mirzaei M, FiroozAbadi M. The impact of data mining on prediction of renal transplantation survival and identifying the effective factors on the transplanted kidney. Journal of Health and Biomedical Informatics. 2016 Jun 10; 3(1):1-9.

19- Moreira LB, Namen AA. A hybrid data mining model for diagnosis of patients with clinical suspicion of dementia. Computer methods and programs in biomedicine. 2018 Oct 1; 165:139-49.

20- Simić S, Banković Z, Simić D, Simić SD. A hybrid clustering approach for diagnosing medical diseases. InInternational Conference on Hybrid Artificial Intelligence Systems 2018 Jun 20 (pp. 741-752). Springer, Cham.

21- Pasanisi S, Paiano R. A hybrid information mining approach for knowledge discovery in cardiovascular disease (CVD). Information. 2018 Apr; 9(4):90.

22- Zhu M, Xia J, Jin X, Yan M, Cai G, Yan J, Ning G. Class weights random forest algorithm for processing class imbalanced medical data. IEEE Access. 2018 Jan 4; 6:4641-52.

Designing an intelligent system for predicting chromosomal genetic diseases using data mining

Fariba Salahi[*1] , Nastarn Farajpour[2]

Submitted: 2021.5.1 Accepted: 2022.1.16

Abstract:

Background and Aim: Today, we are witnessing tremendous advances in medical data mining. The data, by analyzing and discovering the relationships between them, can lead to algorithms that help us prevent or treat many diseases. Meanwhile, genetic diseases have attracted a large part of the attention of the medical world because the birth of children with genetic disorders imposes a great financial, psychological and emotional burden on society. Therefore, the aim of this study was to present an algorithm as a secondary screening test before performing cell and molecular tests.

Material and Methods: In this study, we studied 1000 cases of pregnant women who were in moderate or high-risk group after screening tests. Their clinical data were stored, missing data was deleted, and records were integrated. Then, using Clementine software, data mining and data correlation were performed, and finally a suitable algorithm was performed for diagnosing the disease. Genetic mutations were identified as well.

Results: By applying five algorithms of neural networks, support vector machine, binary decision tree, multiple decision tree and logistic regression on the data, it was found that the neural network algorithm with 97.522% accuracy had the highest success rate in diagnosis of genetic-chromosomal diseases before birth.

Conclusion: The use of genetic algorithm as a screening test causes less people to be candidates for costly and dangerous cellular and molecular tests and can be used as a tool to help detect the disease.

Keywords Classification algorithms, Neural Networks, Chromosomal abnormalities, Support Vector Machine

[1] Assistant Professor, Department of Industrial Management, Faculty of Management, Islamic Azad University, Electronic Branch, Tehran, Iran. (*Corresponding Author) . Email:salahi_en@yahoo.com.
[2] Master of Advanced Information Systems Management, Faculty of Management, Islamic Azad University, Electronic Branch, Tehran, Iran