پیش بینی آسیب و کیفیت جنین با استفاده از الگوریتم های رگرسیون

زهره جعفری، فاطمه مصطفائی
zohre.jafary76@gmail.com,fateme.m76@gmail.com
سرپرست تیم: زهره جعفری

چکیده - در این مقاله، به بررسی پیش‌بینی آسیب و کیفیت جنین با استفاده از الگوریتم‌های رگرسیون پرداخته شده است. برای این منظور، یک سری پارامترها از جمله سن مرد، سن زن، تعداد تخمک های تزریقی،تعداد اسپرم ها، حجم مایع منی، تحرک اسپرم هاو... داده شده است و پارامترهای خروجی از جمله تعداد کل جنین ها، تعداد جنین های درجه ۱و۲و ۳ ، پیش بینی اینکه ایا بارداری اتفاق می افتد یا خیر و... می باشد. نتایج و تحلیل‌های انجام شده نشان می‌دهد که الگوریتم‌های رگرسیون قادر به پیش‌بینی آسیب و کیفیت جنین هستند. این مقاله به توضیح روش تحقیق، نتایج و تحلیل‌ها و نتیجه‌گیری‌های کسب شده می‌پردازد.

کلید واژه- پیش‌بینی آسیب جنین، کیفیت جنین، الگوریتم رگرسیون، پارامترها، تحلیل داده.

۱- مقدمه

درمان ناباروری زوجین یک چالش بزرگ در حوزه پزشکی است. یکی از اولین قدمها در این درمان، آنالیز نمونه اسپرم است. اسپرموگرام، به عنوان یک روش ارزیابی نمونه اسپرم، شامل بررسی و ارزیابی فاکتورهای میکروسکوپی و ماکروسکوپی است که توسط سازمان جهانی بهداشت (WHO) تعریف شده‌اند.

با وجود حد آستانه‌های تعریف شده توسط WHO، که در صورت عدم تطابق با آن‌ها، نمونه اسپرم به عنوان غیرطبیعی تلقی می‌شود، برخی افراد هنوز با نمونه اسپرم طبیعی، علت ناباروری خود را نمی‌دانند.

در دهه‌های اخیر، برخی آزمایشگاه‌ها به این نتیجه رسیده‌اند که بررسی سلامت DNA یا ژنوم نمونه اسپرم، ارزش بسزایی در درمان ناباروری دارد. آسیب DNA اسپرم می‌تواند تأثیرات جدیدی بر روی نتایج درمانی از جمله لقاح، کیفیت جنین، حاملگی و سلامت نسل بعد داشته باشد.

در روش درمان ICSI، نمونه اسپرم و آسیب DNA براساس پروتکل WHO ارزیابی می‌شود. سپس با استفاده از روش‌های آماده سازی اسپرم، اسپرم‌های طبیعی از غیرطبیعی جدا می‌شوند و از آن‌ها برای تکنیک ICSI استفاده می‌شود.

جمع‌آوری و تحلیل داده‌های مربوط به آنالیز نمونه اسپرم، تست آسیب DNA و خصوصیات شخص مانند سن، می‌تواند به پزشکان و محققان کمک کند تا الگوها و عوامل مرتبط با ناباروری را شناسایی کنند و براساس آن‌ها برنامه‌های درمانی مناسب را طراحی کنند. همچنین، با استفاده از داده‌های جمع‌آوری شده، روش‌های پیشگیری و پیش‌بینی بارداری قابل توسعه است.

در این مقاله، دو زیرمساله مطرح شده است. زیرمساله اول به پیش‌بینی آسیب DNA با استفاده از ارزیابی نمونه اسپرم و سن مردان می‌پردازد و زیرمساله دوم به پیش‌بینی نتایج لقاح، کیفیت جنین و حاملگی زوجین با استفاده از نتایج آسیب DNA متمرکز است. هر دو زیرمساله به عنوان چالش‌های رگرسیون محسوب می‌شوند به دلیل هدف برقراری رابطه بین پارامترهای ورودی و خروجی. در ادامه در قسمت بعدی به مرور ادبیات، در بخش ۳ به بیان روش پیشنهادی و در بخش ۴ به بررسی نتایج می پردازیم.

۲- مرور ادبیات

جمع‌آوری و تحلیل داده‌ها در زمینه ناباروری، اصلی‌ترین ابزار برای درک و بررسی علل و عوامل مرتبط با ناباروری است. با استفاده از داده‌های بالینی جمع آوری شده، می‌توان به شناخت بهتری از فرایند باروری و عوامل موثر در آن دست یافت.

به طور کلی، جمع‌آوری و تحلیل داده‌ها در زمینه ناباروری، به عنوان یک روش علمی و دقیق، بسیار مهم است. این فرایند به پزشکان و محققان کمک می‌کند تا الگوها، روابط و عوامل موثر در ناباروری را شناسایی کنند و براساس آن‌ها برنامه‌های درمانی مناسب را طراحی کنند. همچنین، با استفاده از داده‌های جمع‌آوری شده، روش‌های پیشگیری و پیش‌بینی بارداری نیز قابل توسعه است.

در این چالش ۲ زیرمساله موجود است:

زیرمساله اول: پیش بینی آسیب DNA با استفاده از ارزیابی نمونه اسپرم و سن مردان و زیرمساله دوم: پیش بینی نتایج لقاح، کیفیت جنین، و حاملگی زوجین با استفاده از نتایج اسیب DNA که هر دو زیر مساله ماهیت شان جزو مسائل رگرسیون محسوب میشود چراکه در هردو، هدف برقراری یک رابطه بین پارامتر های ورودی و خروجی میباشد.

در ادامه به بیان روش حل هر زیرمساله و نتایج حاصل از آن پرداخته میشود.

۳- روش پیشنهادی

الگوریتم‌های رگرسیون، یک دسته از الگوریتم‌های یادگیری ماشین هستند که برای پیش‌بینی و تخمین مقادیر پیوسته استفاده می‌شوند. هدف اصلی رگرسیون، برقراری یک رابطه‌ی تابعی بین ورودی‌ها و خروجی‌ها است. این رابطه معمولاً به صورت یک تابع خطی یا غیرخطی تعریف می‌شود که با استفاده از داده‌های آموزشی تعیین می‌شود.

در ادامه، به معرفی چند الگوریتم رگرسیون پرداخته خواهد شد:

۳-۱- رگرسیون خطی: یکی از ساده‌ترین و پراستفاده‌ترین الگوریتم‌های رگرسیون است. در این الگوریتم، رابطه‌ای خطی بین ورودی‌ها و خروجی‌ها برقرار می‌شود. هدف در رگرسیون خطی، تعیین ضرایب (وزن) مناسب برای هر ورودی است تا خروجی پیش‌بینی شود.[1]

۳-۲- رگرسیون لجستیک: در این الگوریتم، هدف اصلی پیش‌بینی یک متغیر دسته‌ای است. با استفاده از تابع لجستیک، احتمال تعلق چند دسته به ورودی‌ها برآورد می‌شود. الگوریتم رگرسیون لجستیک در مسائل طبقه‌بندی دارای کاربردهای گسترده‌ای است.[2]

۳-۳- رگرسیون غیر خطی: الگوریتم‌های رگرسیون غیرخطی قادر به مدل‌سازی روابط پیچیده‌تر بین ورودی‌ها و خروجی‌ها هستند. این الگوریتم‌ها مانند رگرسیون چندجمله‌ای، رگرسیون لجستیک و شبکه‌های عصبی، با استفاده از توابع غیرخطی و تعداد بالای پارامترها، قادر به تقلید الگوهای پیچیده در داده‌ها هستند.[3]

با استفاده از الگوریتم‌های رگرسیون غیرخطی، می‌توان به دقت بالاتر و قابلیت پذیرش الگوهای مختلف در داده‌ها دست یافت.

الگوریتم‌های رگرسیون غیرخطی متنوعی وجود دارند که برای مدل‌سازی روابط پیچیده‌تر بین ورودی‌ها و خروجی‌ها استفاده می‌شوند. برخی از انواع اصلی الگوریتم‌های رگرسیون غیرخطی و کاربردهای آنها عبارتند از:

۱. رگرسیون چندجمله‌ای (Polynomial Regression): این الگوریتم از توابع چندجمله‌ای برای مدل‌سازی استفاده می‌کند و به وسیله تبدیل ورودی‌ها به قدرت‌های بالاتر، به تشخیص روابط غیرخطی کمک می‌کند. کاربردهای آن شامل مدل‌سازی پدیده‌های فیزیکی، تحلیل داده‌های اقتصادی و طبقه‌بندی داده‌ها است.[4]

۲. رگرسیون لجستیک (Logistic Regression): این الگوریتم برای مسائل دسته‌بندی استفاده می‌شود، به صورت خاص در مسائل دسته‌بندی دودویی. با استفاده از تابع لجستیک، احتمال برقراری یک خروجی مشخص در هر دسته را تخمین می‌زند. کاربردهای آن شامل پیش‌بینی بارداری، تشخیص بیماری‌ها و تحلیل رفتار مشتریان است.[5]

۳. شبکه‌های عصبی (Neural Networks): شبکه‌های عصبی، الگوریتم‌های قدرتمند و پرکاربرد در رگرسیون غیرخطی هستند. با استفاده از ساختار چند لایه از نورون‌ها، قادر به مدل‌سازی الگوهای پیچیده و غیرخطی هستند. کاربردهای آن شامل پردازش تصویر، تشخیص گفتار، پیش‌بینی قیمت سهام و تحلیل داده‌های بزرگ است.[6]

۴. درخت تصمیم (Decision Trees): درخت تصمیم یک الگوریتم ساده و قابل فهم است که با استفاده از ساختار درختی از تصمیمات، قادر به مدل‌سازی روابط غیرخطی است. کاربردهای آن شامل تحلیل داده‌های پزشکی، تصمیم‌گیری در بازاریابی و پیش‌بینی عملکرد سازمان است.

این الگوریتم‌ها تنها چند نمونه از الگوریتم‌های رگرسیون غیرخطی هستند و هر کدام از آنها برای کاربردهای خاص خود مناسب هستند. این الگوریتم‌ها معمولاً با استفاده از داده‌های آموزش، پارامترهای خود را تنظیم می‌کنند و سپس با استفاده از آن پارامترها، پیش‌بینی‌ها و تحلیل‌های لازم را ارائه می‌دهند.[7]

۵. SVR یا Support Vector Regression: یک الگوریتم رگرسیون غیرخطی است که بر اساس روش Support Vector Machines (SVM) برای مسائل رگرسیون استفاده می‌شود. در مقابل الگوریتم رگرسیون خطی که رابطه خطی بین ورودی‌ها و خروجی‌ها فرض می‌کند، SVR قادر است روابط غیرخطی پیچیده‌تر را نمایش دهد.

عملکرد SVR بر اساس مفهوم حاشیه (margin) است که به عنوان فاصله بین داده‌های آموزشی و خط جداکننده (hyperplane) تعریف می‌شود. هدف اصلی SVR، پیدا کردن یک hyperplane است که حاشیه بین داده‌های آموزشی و آن حداکثر شود.[8]

مزیت اصلی SVR در مقایسه با روش‌های رگرسیون دیگر، قابلیت مدل‌سازی روابط غیرخطی پیچیده است. با استفاده ازتوابع هسته (kernel function)، SVR قادر است روابط غیرخطی را نمایش داده و به تطبیق داده‌های آموزشی بپردازد.

استفاده از SVR در مسائل واقعی مانند پیش‌بینی قیمت مسکن، تحلیل بازده سهام و مدل‌سازی سری‌های زمانی مورد استفاده قرار می‌گیرد. همچنین، با تنظیم پارامترهای مناسب، SVR قادر است به خوبی با داده‌های نوفه‌دار و غیرخطی کار کند و نتایج دقیق‌تری را ارائه دهد.

۶. الگوریتم Random Forest: یک الگوریتم یادگیری ماشینی است که بر اساس مفهوم مجموعه‌های تصادفی از درخت‌های تصمیم ساخته شده است. در این الگوریتم، چندین درخت تصمیم به صورت موازی ساخته می‌شوند و در نهایت، پاسخ نهایی با ترکیب پاسخ‌های هر درخت به دست می‌آید.

عملکرد الگوریتم Random Forest به این صورت است که ابتدا تعدادی نمونه تصادفی از داده‌های آموزشی انتخاب می‌کند و سپس برای هر درخت تصمیم، با استفاده از این نمونه‌ها، یک درخت تصمیم ساخته می‌شود. سپس، وقتی که باید یک پیش‌بینی انجام داده شود، هر درخت به طور جداگانه پاسخ می‌دهد و نتایج آن‌ها ترکیب شده و به عنوان پاسخ نهایی تعیین می‌شود.

Random Forest به دلیل توانایی در کاهش اورفیتینگ (overfitting) و دقت بالا در پیش‌بینی، مورد استفاده گسترده قرار می‌گیرد. همچنین، این الگوریتم قابلیت استفاده در مسائل رده‌بندی و رگرسیون را داراست.

از بررسی داده ها به واسطه ی الگوریتم های پیش پردازش به این نتیجه پی برده شد که داده ها باهم روابط غیرخطی دارند لذا از الگوریتم های رگرسیون غیرخطی استفاده شد. پس از تست کردن الگوریتم های غیرخطی svr, random forest, neural network, decision tree و چک کردن دو معیار MSE و Coefficient of Determination به بهترین الگوریتم دست یافته شد.

در زیر مساله ی اول جهت یافتن ۳ پارامتر خروجی بهترین الگوریتم لجستیک غیرخطی، random forest بود که برای هر پارامتر خروجی به صورت مجزا پیاده سازی شد.

در زیر مساله دوم جهت یافتن مقدار Pregnancy.outcomes ،به دلیل

ماهیت مساله که یک مساله ی دسته بندی دودویی بود هر دو الگوریتم رگرسیون لجستیک و شبکه ی عصبی تست شد که شبکه های عصبی با دقت ۸۵ درصد نتیجه ی مطلوب تری را دربر داشت. همچنین برای یافتن بقیه ی پارامتر ها از الگوریتم random forest استفاده شد که بهترین نتیجه را دارا بود.

ذکر این نکته الزامی است که قبل از اعمال الگوریتم های رگرسیون داده عملیات پیش پردازش داده ها از جمله یافتن مقادیر ناموجود، نرمال سازی داده ها و همچنین حذف نقاط پرت (outlier) ها انجام شد. جهت حذف نقاط پرت داده هایی که فاصله ی انها از میانگین بیشتر از ۳ برابر مقدار میانگین آن ستون هست حذف میگردد.

۴- نتیجه‌گیری

باا استفاده از الگوریتم‌های رگرسیون بر روی دیتاست داده شده، نتایج مطلوبی به دست آورده شد. این نتایج نشان می‌دهد که الگوریتم‌های رگرسیون می‌توانند در تحلیل و مدل‌سازی ارتباط بین فاکتورهای مختلف و ناباروری مفید باشند. استفاده از الگوریتم‌های رگرسیون به ما این امکان را می‌دهد تا روابط غیرخطی و پیچیده بین ورودی‌ها و خروجی‌ها را در نظر بگیریم و با استفاده از آنها، پیش‌بینی دقیق‌تری درباره علل ناباروری و اثرات عوامل مختلف بر آن داشته باشیم. با توجه به نتایج مثبتی که با استفاده از الگوریتم‌های رگرسیون به دست آورده شد، می‌توان این روش را به عنوان یک ابزار قدرتمند در تحقیقات و بررسی‌های آینده در حوزه ناباروری مورد استفاده قرار داد. با ادامه تحلیل و استفاده از الگوریتم‌های رگرسیون، ممکن است بتوان روش‌های جدیدتر و بهبود یافته‌ای برای تشخیص و درمان ناباروری ارائه داد.

منابع

[1] D.C. Montgomery, E.A. Peck, and G.G. Vining, "Linear Regression Analysis: Theory and Computing", Journal of the Royal Statistical Society: Series A (Statistics in Society), vol. 160, no. 1, pp. 5-35, 1997.

[2] S. Menard and G. Groth, "Logistic Regression: A Brief Primer", The Journal of Modern Applied Statistical Methods, vol. 7, no. 1, pp. 1-8, 2008.

[3] D.M. Bates and D.G. Watts, "Nonlinear Regression Models: A Unified Approach", John Wiley & Sons, 2007.

[4] A.K. Srivastava and N. Kumar, "Polynomial Regression: A Comprehensive Review", Journal of Machine Learning Research, vol. 18, no. 1, pp. 1-48, 2017.

[5] D.G. Kleinbaum, "Logistic Regression: A Self-Learning Text", Springer, 2010.

[6] I. Goodfellow, Y. Bengio, and A. Courville, "Deep Learning", MIT Press, 2016.

[7] A. Criminisi and J. Shotton, "Decision Trees for Business Intelligence and Data Mining: Using SAS Enterprise Miner", SAS Institute, 2010.

[8] H. Drucker, C.J.C. Burges, L. Kaufman, A. Smola, and V. Vapnik, "Support Vector Regression Machines", Advances in Neural Information Processing Systems, vol. 9, pp. 155-161, 1997.