دانشگاه خوارزمی

دانشکده علوم ریاضی و کامپیوتر

عنوان پروژه: شناسایی و طبقه‌بندی سرطان سینه با انتخاب ویژگی رگرسیون لجستیک و طبقه‌بند GMDH

استاد و سرپرست:

دکتر کیوان برنا

دانشجو:

فاطمه خدادادی

زمستان1401

فهرست مطالب

فهرست اشکال

فهرست جداول

1- مقدمه

سرطان سینه یک بیماری بسیار شایع بین بانوان است. این سرطان به دلیل رشد غیرعادی سلول‌ها در سینه و به دنبال آن ایجاد تومور شکل می‌گیرد. تومور سرطان سینه می‌تواند خوش‌خیم یا بدخیم باشد. در سنین 25 تا 30 سال به‌ندرت سرطان سینه مشاهده می‌شود و حدود 25% از مرگ‌ومیر ناشی از سرطان سینه بین بانوان 40 تا 49 است. بنا بر گزارش‌های سازمان بهداشت جهانی، از هر 8 تا 10 زن یک نفر از سرطان سینه رنج می‌برد و تحقیقات نشان می‌دهد که علت نرخ مرگ‌ومیر بالا ناشی از سرطان سینه تشخیص دیر هنگام آن است.

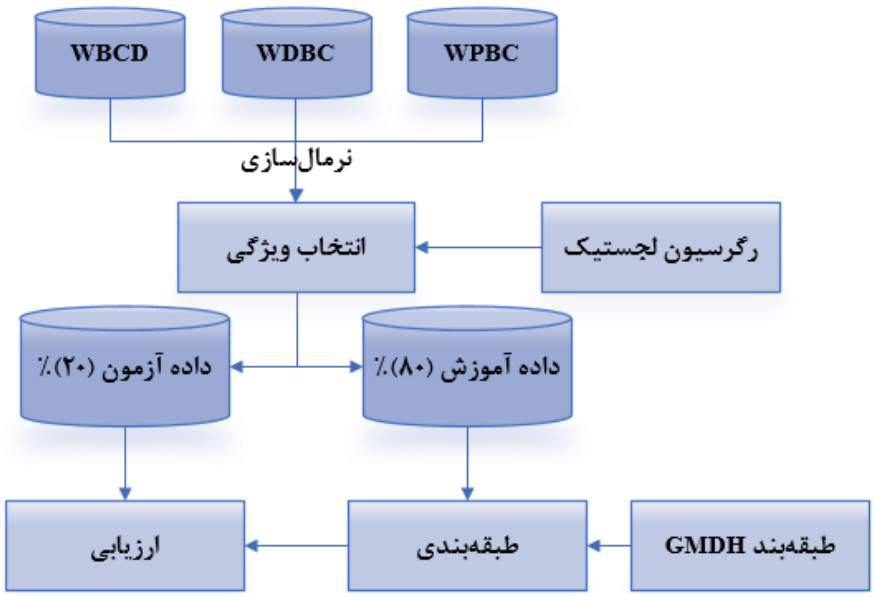
روش‌های تشخیص سرطان سینه مبتنی بر رایانه معمولاً بیماران را به دو گروه خوش‌خیم (بدون سرطان) و بدخیم (با سرطان) طبقه‌بندی می‌کنند. طیف وسیعی از این روش‌ها وجود دارند که برخی طبقه‌بندی را با انتخاب ویژگی (متغیر مستقل) انجام می‌دهند و برخی دیگر بدون انتخاب ویژگی در این پروژه برای حذف ویژگی‌های نا مؤثر، با رگرسیون لجستیک پیاده‌سازی می‌گردد. به کمک رگرسیون لجستیک که یک تابع خطی و کم‌هزینه است، به شکل مؤثری می‌توان انتخاب ویژگی را انجام داد. در گام بعدی الگوریتم طبقه‌بندی GMDH بر پایگاه داده اعمال می‌گردد.

2- روش‌شناسی

مطابق با شکل 1، روش‌شناسی پروژه شامل مراحل زیر است:

1. ابتدا مجموعه داده‌های WBCD، WDBC و WPBC جمع‌آوری و ویژگی‌های داده (متغیر های مستقل) آن‌ها با استفاده از رابطه نرمال‌سازی خطی، در بازه 0-1 نرمال‌سازی می‌شود. سپس با برازش رگرسیون لجستیک مهم‌ترین ویژگی‌ها برای طبقه‌بندی انتخاب می‌گردند.

2. به‌صورت تصادفی 80% از نمونه‌ها برای آموزش و 20% باقی نیز برای ارزیابی طبقه‌بندی در نظر گرفته می‌شود.
3. طبقه‌بندی GMDH با داده آموزش ایجاد می‌گردد و عملکرد آن با داده آزمون مورد ارزیابی قرار می‌گیرد.



شکل 1: روش‌شناسی

3- پایگاه داده

در این پروژه از سه مجموعه داده WBCD، WDBC و WPBC استفاده می‌شود. این مجموعه داده‌ها از Diagnostic Wisconsin Breast Cancer Database [1] جمع‌آوری‌شده‌اند. مجموعه داده WBCD شامل 699 نمونه (458 خوش‌خیم و 241 بدخیم) و نه ویژگی تشکیل‌شده است.

به همین ترتیب مجموعه داده WDBC ترتیب از 569 (356 خوش‌خیم و 213 بدخیم) نمونه داده و 30 ویژگی و مجموعه داده WPBC ترتیب از 194 (148 خوش‌خیم و 46 بدخیم) نمونه داده و 33 ویژگی تشکیل‌شده‌اند.

4- انتخاب ویژگی رگرسیون لجستیک

نتایج انتخاب ویژگی رگرسیون لجستیک در جدول 1 خلاصه‌شده است. مقایسه نتایج نشان می‌دهد که ضرایب به‌دست‌آمده با ضرایب مقاله مرجع [2] یکسان نیستند. بااین‌وجود ضرایب محاسبه‌شده به‌صورت نسبی بسیار مشابه با نتایج مقاله هستند. به‌عنوان‌مثال مقاله ویژگی‌های 1، 3، 6، 7 و 9 را از مجموعه داده WBCD انتخاب کرده است و این پروژه نیز به اهمیت بالای ویژگی‌های 1، 3، 6 و 7 رسیده است. از میان 15 ویژگی که مقاله از مجموعه داده WDBC انتخاب کرد، 11 ویژگی دارای بالاترین اهمیت طبق نتایج پروژه هستند. به همین صورت در مجموعه داده WPBC مقاله مجموعاً 16 ویژگی را مهم ارزیابی کرد که از میان آن‌ها پروژه برای نه ویژگی به ضرایب بالا رسیده است.

با توجه به جدول 1، برای مجموعه داده WBCD ویژگی‌های شماره 6، 1، 7، 3 و 8، برای مجموعه داده WDBC ویژگی‌های شماره 28، 21، 22، 23، 8، 1، 3، 24، 2، 4، 25، 7، 27، 29 و 11 و برای مجموعه داده WPBC ویژگی‌های شماره 26، 33، 22، 32، 24، 6، 25، 14، 17، 21، 12، 31، 20، 29 و 28 برای طبقه‌بندی انتخاب شدند.

جدول 1: نتایج انتخاب ویژگی رگرسیون لجستیک

مجموعه داده WPBC		مجموعه داده WDBC		مجموعه داده WBCD	
شماره ویژگی	ضریب ویژگی	شماره ویژگی	ضریب ویژگی	شماره ویژگی	ضریب ویژگی
1	-2.42279	1	1.894344	1	2.736719
2	-0.1151	2	1.716611	2	1.543698
3	-0.66001	3	1.858307	3	1.91172
4	-0.0879	4	1.603036	4	1.563307
5	-0.00212	5	0.65249	5	1.011442
6	0.538746	6	0.334192	6	2.779279
7	-0.26622	7	1.436259	7	1.926249
8	-0.31697	8	2.14063	8	1.566187
9	0.009056	9	0.548293	9	1.146398
10	-0.4318	10	-0.97916		

ادامه جدول 1: نتایج انتخاب ویژگی رگرسیون لجستیک

مجموعه داده WPBC		مجموعه داده WBCD	
شماره ویژگی	شماره ویژگی	ضریب ویژگی	شماره ویژگی
-0.37034	11	1.284867	11
0.217122	12	0.038806	12
-0.68465	13	0.989682	13
0.372891	14	0.849395	14
0.024979	15	0.059908	15
-0.07709	16	-0.65866	16
0.266786	17	-0.27117	17
-0.60252	18	0.269023	18
-0.76097	19	-0.22802	19
0.133756	20	-0.63954	20
0.222405	21	2.433212	21
0.611704	22	2.359974	22
-0.39996	23	2.223758	23
0.585324	24	1.75955	24
0.517069	25	1.577173	25
0.847839	26	0.777522	26
-0.01758	27	1.383559	27
0.026002	28	2.730713	28
0.070108	29	1.337405	29
0.01635	30	0.335405	30
0.19809	31		
0.608063	32		
0.822759	33		

5- طبقه‌بندی با الگوریتم GMDH

پس از آموزش الگوریتم GMDH، ارزیابی طبقه‌بندی با معیارهای AUC، Precision، Recall و F1 score و ماتریس درهم‌ریختگی با داده آموزش و ارزیابی انجام شد. جداول 2 و 3 به ترتیب نتایج ارزیابی طبقه‌بندی با داده آموزش و آزمون را نشان می‌دهند و ماتریس‌های درهم‌ریختگی نیز در اشکال 2، 3 و 4 آمده است. همچنین مقایسه بین معیارهای ارزیابی در حالت انتخاب ویژگی پیش از طبقه‌بندی و عدم انتخاب ویژگی پیش از طبقه‌بندی با داده آموزش و آزمون به ترتیب در اشکال 5 و 6 آمده است.

نتایج مقایسه نتایج طبقه‌بندی با مقاله مرجع به شرح زیر است:

1. مطابق با مقاله مرجع، بهترین عملکرد طبقه‌بندی به ترتیب با مجموعه داده‌های WDBC، WBCD و WPBC مشاهده شد.
2. مطابق با مقاله مرجع، ارزیابی طبقه‌بندی مجموعه داده‌های WDBC و WBCD با AUC، Precision، Recall و F1 score مقادیر بالا و نزدیک به یک را نشان داد.
3. برخلاف مقاله مرجع، در این پروژه طبقه‌بندی مجموعه داده WPBC با عملکرد ضعیف انجام شد.
4. هم‌راستا با مقاله مرجع (شکل 5 و 6) نتیجه‌گیری شد که انتخاب ویژگی با رگرسیون لجستیک در بهبود عملکرد طبقه‌بندی بسیار مؤثر بوده است.
5. علت تفاوت‌های میان مشاهدات این پروژه و مقاله مرجع را در: 1- مقادیر متفاوت پارامترهای الگوریتم‌های رگرسیون لجستیک و GMDH و 2- توزیع متفاوت نمونه‌های آموزش و ارزیابی می‌توان دانست.

جدول 2: نتایج ارزیابی طبقه‌بندی با داده آموزش

	AUC	Precision	Recall	F1 score
طبقه‌بندی با مجموعه داده WBCD	0.958691	0.96875	0.934673	0.951407
طبقه‌بندی با مجموعه داده WDBC	0.953101	0.9875	0.913295	0.948949
طبقه‌بندی با مجموعه داده WPBC	0.617678	0.785714	0.261905	0.392857

جدول 3: نتایج ارزیابی طبقه‌بندی با داده آزمون

	AUC	Precision	Recall	F1 score
طبقه‌بندی با مجموعه داده WBCD	0.961727	0.886364	0.975	0.928571
طبقه‌بندی با مجموعه داده WDBC	0.974359	1	0.948718	0.973684
طبقه‌بندی با مجموعه داده WPBC	0.582143	0.25	0.25	0.25

	مثبت واقعی	منفی واقعی
مثبت پیش بینی شده	342	11
منفی پیش بینی شده	13	180

	مثبت واقعی	منفی واقعی
مثبت پیش بینی شده	90	1
منفی پیش بینی شده	5pai	41

شکل 2: ماتریس درهم‌ریختگی طبقه‌بندی با مجموعه داده WBCD با داده آموزش (چپ) و آزمون (راست)

		منفی واقعی	مثبت واقعی
مثبت پیش بینی شده		0	65
منفی پیش بینی شده		46	3

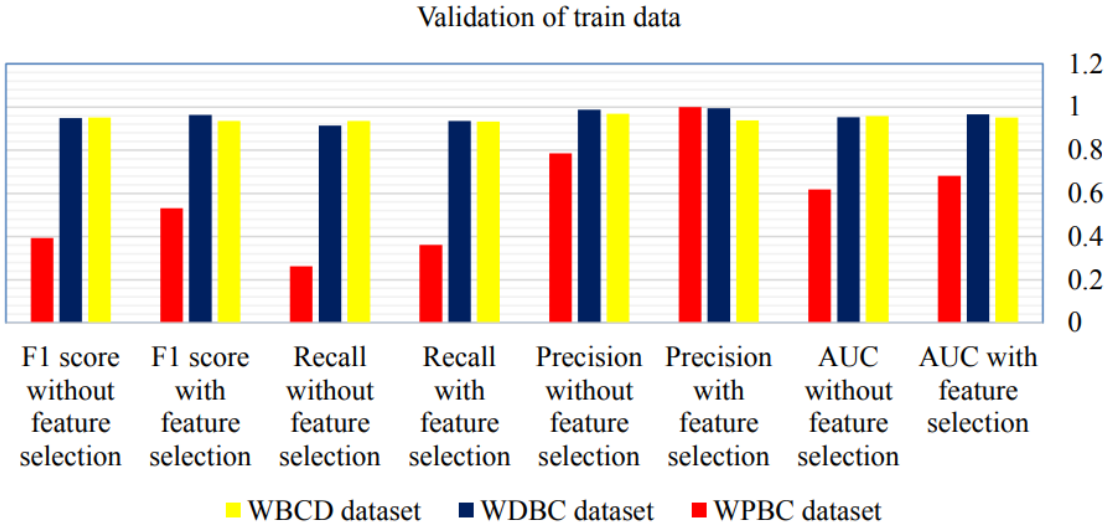
		منفی واقعی	مثبت واقعی
مثبت پیش بینی شده		0	292
منفی پیش بینی شده		150	13

شکل 3: ماتریس درهم‌ریختگی طبقه‌بندی با مجموعه داده WDBC با داده آموزش (چپ) و آزمون (راست)

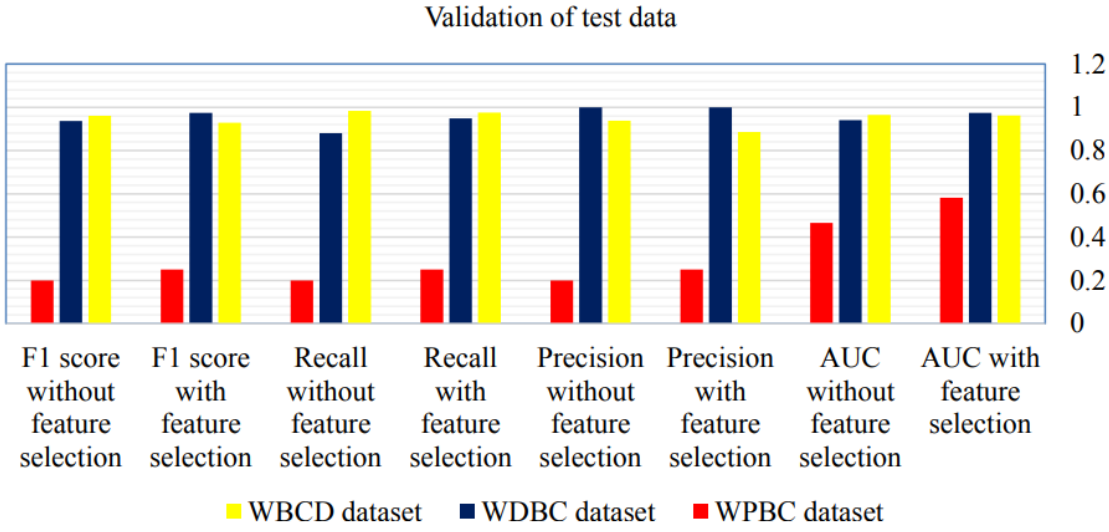
		منفی واقعی	مثبت واقعی
مثبت پیش بینی شده		3	32
منفی پیش بینی شده		1	3

		منفی واقعی	مثبت واقعی
مثبت پیش بینی شده		3	110
منفی پیش بینی شده		11	31

شکل 4: ماتریس درهم‌ریختگی طبقه‌بندی با مجموعه داده WPBC با داده آموزش (چپ) و آزمون (راست)



شکل 5: مقایسه بین معیارهای ارزیابی با انتخاب ویژگی و بدون انتخاب ویژگی پیش از طبقه‌بندی با داده آموزش



شکل 6: مقایسه بین معیارهای ارزیابی با انتخاب ویژگی و بدون انتخاب ویژگی پیش از طبقه‌بندی با داده آزمون

مراجع

[1] UCI. "Breast Cancer Wisconsin (Original) Data Set." https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original) (accessed 2023/23/1).

[2] Z. Khandezamin, M. Naderan, and M. J. Rashti, "Detection and classification of breast cancer using logistic regression feature selection and GMDH classifier," *Journal of Biomedical Informatics*, vol. 111, p. 103591, 2020.