پروژه پایانی درس یادگیری ماشین-پیش بینی بیماری های قلبی با استفاده از یادگیری ماشین

آزیتا شیخی
دانشگاه خوارزمی
کارشناسی ارشد علوم کامپیوتر
گرایش علوم تصمیم و دانش

بهمن ۱۴۰۱

۱ چکیده

در زمانهای اخیر، پیشبیني بیماریهای قلبي یکي از پیچیده ترین کارها در حوزه پزشکي بوده است. در عصر مدرن، تقریباً یک نفر در دقیقه به دلیل بیماری قلبي جان خود را از دست مي دهد. علم داده نقش مهمي در پردازش حجم عظیمي از داده ها در حوزه مراقبت های بهداشتي ایفا مي کند. از آنجایي که پیشبیني بیماری قلبي یک کار پیچیده است، نیاز به خودکارسازی فرآیند پیشبیني برای جلوگیری از خطرات مرتبط با آن و آگاه کردن بیمار از قبل وجود دارد. این مقاله از مجموعه داده های بیماری قلبی موجود در مخزن یادگیری ماشیني UCI استفاده میکند. کار پیشنهادی شانس بیماری قلبی را پیش بیني میکند و سطح خطر بیمار را با اجرای تکنیکهای مختلف داده کاوی مانند Random Forest, Decision Tree, Naive Bayes, Logistic Regression طبقه بندی میکند. بنابراین، این مقاله یک مطالعه مقایسه ای را با تجزیه و تحلیل عملکرد الگوریتم های مختلف یادگیری ماشین ارائه مي دهد. نتایج آزمایشي تأیید مي کند که الگوریتم جنگل تصادفي بالاترین دقت 90.16% و بیشترین پیش بینی درست 94.11% برای افرادی که احتمال بیماری قلبی دارند را در مقایسه با سایر الگوریتم های پیاده سازی شده بدست آورده است.

۲ نکاتی درباره دیتاست

۱.دیتاست مقاله در فایل گزارش با نام heart.csv موجود میباشد.
۲.دیتاست سطرهای تکراری ندارد.
۳.دیتاست مقادیر null ندارد.

۴.مقادیر همه ی ستون های دیتاست عددی هستند.
۵.این دیتاست را به ۸۰ درصد train و ۲۰ درصد test تبدیل میکنیم.
۶.train این دیتاست دارای ۱۱ نمونه با حداقل یک مقدار پرت هست که به علت تعداد نمونه های کم دیتاست، هیچکدام از آنهارا حذف نمیکنیم.با حذف آنها نیز دقت برای بعضی معیارها افزایش و برای بعضی دیگر کاهش یافت.
۷.میتوانیم داده ها را scale کنیم و محدوده داده ها را بین ۰ و ۱ درنظر بگیریم.ولی چون دقت برای مدل ها کاهش می یافت،این کار را انجام ندادیم.

۳ نتایج بدست آمده از مدل ها طبق مقاله

Algorithm	True Positive	False Positive	False Negative	True Negative
Logistic Regression	22	5	4	30
Naive Bayes	21	6	3	31
Random Forest	22	5	6	28
Decision Tree	25	2	4	30

شکل ۱: ماتریس کانفیوژن بدست آمده از مدل ها طبق مقاله

Algorithm	Precision	Recall	F-measure	Accuracy
Decision Tree	0.845	0.823	0.835	81.97%
Logistic Regression	0.857	0.882	0.869	85.25%
Random Forest	0.937	0.882	0.909	90.16%
Naive Bayes	0.837	0.911	0.873	85.25%

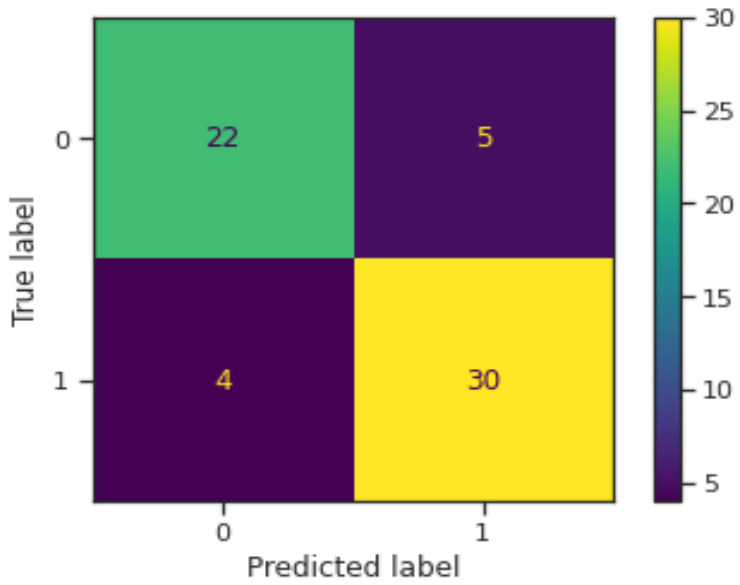
شکل ۲: آنالیز الگوریتم ها طبق مقاله

۴ نتایج بدست آمده از مدل ها طبق پیاده سازیِ انجام شده

۱.۴ مدل Logistic Regression

نتایج رگرسیون لجستیک بر روی این دیتاست به صورت زیر میباشد:

Precision: 0.857
Recall: 0.882
F1 Score: 0.870
Accuracy: 85.25%



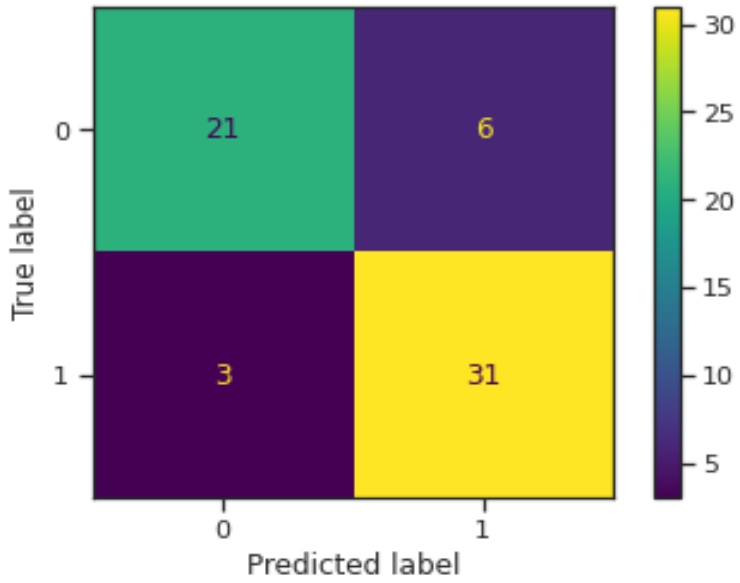
شکل ۳: ماتریس کانفیوژن مدل رگرسیون لجستیک

در این مسئله پیش بینی درست برای افرادی که احتمال بیماری قلبی دارند بسیار بسیار پر اهمیت است و میتوان گفت که اشکالی ندارد اگر کسی که احتمال مبتلا به بیماری قلبی ندارد را اشتباه پیش بینی کنیم ولی حتما باید کسی که احتمال ابتلا به بیماری قلبی دارد را تشخیص بدهیم و درست پیش بینی بکنیم تا بتوانیم از بیماری او پیشگیری کنیم که باعث مرگ و میر و زیان های جبران ناپذیر نشود. همانطور که مشخص است، این مدل ۴ نفر از ۳۴ نفری که احتمال بیماری قلبی دارند را اشتباه پیش بینی میکند(تقریبا ۱۱% پیش بینی نادرست برای افرادی که احتمال بیماری قلبی دارند) که همانطور که گفتیم برای ما پیش بینی درست برای افرادی که احتمال بیماری قلبی دارند بسیار پراهمیت است و این مدل تقریبا ۸۸% مواقع افرادی که احتمال بیماری قلبی دارند را درست تشخیص میدهد.
همچنین نتایج بدست آمده از پیاده سازی این مدل با نتایج بدست آمده در مقاله (طبق شکل ۱ و ۲)یکسان هستند.

۲.۴ مدل Naive Bayes

نتایج بیز ساده بر روی این دیتاست به صورت زیر میباشد:

Precision: 0.838
Recall: 0.912
F1 Score: 0.873
Accuracy: 85.25%



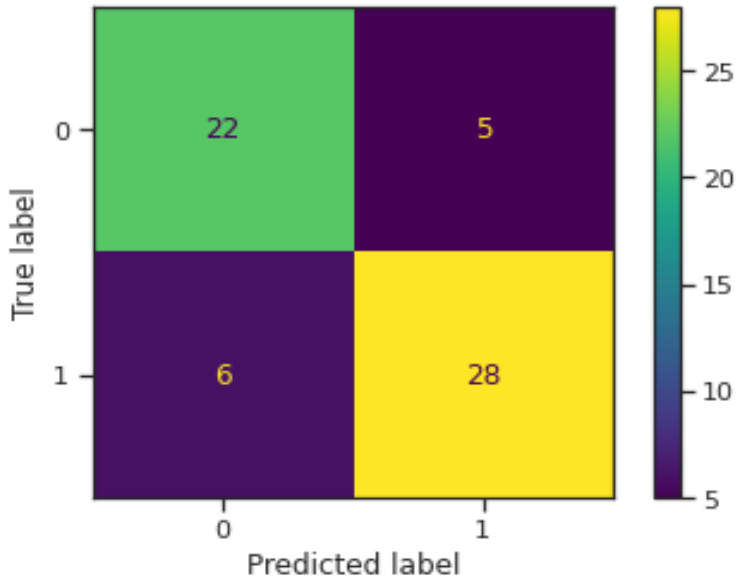
شکل ۴: ماتریس کانفیوژن مدل بیز ساده

همانطور که مشخص است این مدل، ۳ نفر از ۳۴ نفری که احتمال بیماری قلبی دارند را اشتباه پیش بینی میکند (تقریبا ۸% پیش بینی نادرست برای افرادی که احتمال بیماری قلبی دارند) که همانطور که گفتیم برای ما پیش بینی درست برای افرادی که احتمال بیماری قلبی دارند بسیار پراهمیت است و این مدل تقریبا ۹۱% مواقع، افرادی که احتمال بیماری قلبی دارند را درست تشخیص میدهد و این مدل برای این مسئله نسبت به مدل رگرسیون لجستیک عملکرد بهتری دارد.
همچنین نتایج بدست آمده از پیاده سازی این مدل با نتایج بدست آمده در مقاله(طبق شکل ۲) تفاوت بسیار ناچیزی دارند و میتوان گفت نتایج هر دو یکسان هستند و ماتریس آشفتگی در هر دو نتیجه یکی است.

۳.۴ مدل Decision Tree

نتایج درخت تصمیم بر روی این دیتاست به صورت زیر میباشد:

Precision: 0.8485
Recall: 0.8235
F1 Score: 0.8358
Accuracy: 81.97%



شکل ۵: ماتریس کانفیوژن مدل درخت تصمیم

همانطور که مشخص است،این مدل ۶ نفر از ۳۴ نفری که احتمال بیماری قلبی دارند را اشتباه پیش بینی میکند (تقریبا ۱۷% پیش بینی نادرست برای افرادی که احتمال میرود بیماری قلبی داشته باشند) و این مدل تقریبا ۸۲% مواقع افرادی که بیماری قلبی دارند را تشخیص میدهد که همانطور که مشخص است این مدل برای این مسئله نسبت به هر دو مدل رگرسیون لجستیک و بیز ساده عملکرد بدتری دارد. تاکنون مدل ها از نظر عملکر ترتیب زیر را دارند:

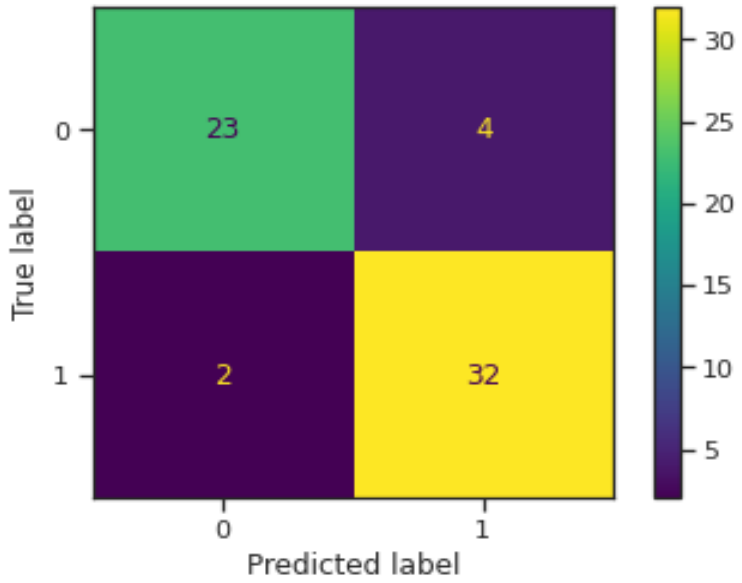
$$\text{Naive Bayes} > \text{Logistic Regression} > \text{Decision Tree}$$

همچنین درخت تصمیم پیاده سازی شده در مقایسه با مقاله (شکل ۱ و ۲) نتایج مشابهی داشتند.

۴.۴ مدل Random Forest

نتایج جنگل تصادفی بر روی این دیتاست به صورت زیر میباشد:

Precision: 0.941
Recall: 0.888
F1 Score: 0.914
Accuracy: 90.16%



شکل ۶: ماتریس کانفیوژن مدل جنگل تصادفی

همانطور که مشخص است،این مدل ۲ نفر از ۳۴ نفری که که احتمال بیماری قلبی دارند را اشتباه پیش بینی میکند(تقریبا ۵% پیش بینی نادرست برای افرادی که احتمال بیماری قلبی دارند) و این مدل تقریبا ۹۴% مواقع افرادی که احتمال بیماری قلبی دارند را درست تشخیص میدهد و همانطور که مشخص است این مدل برای این مسئله نسبت به هر سه مدل رگرسیون لجستیک، بیز ساده و درخت تصمیم عملکرد بهتری دارد.

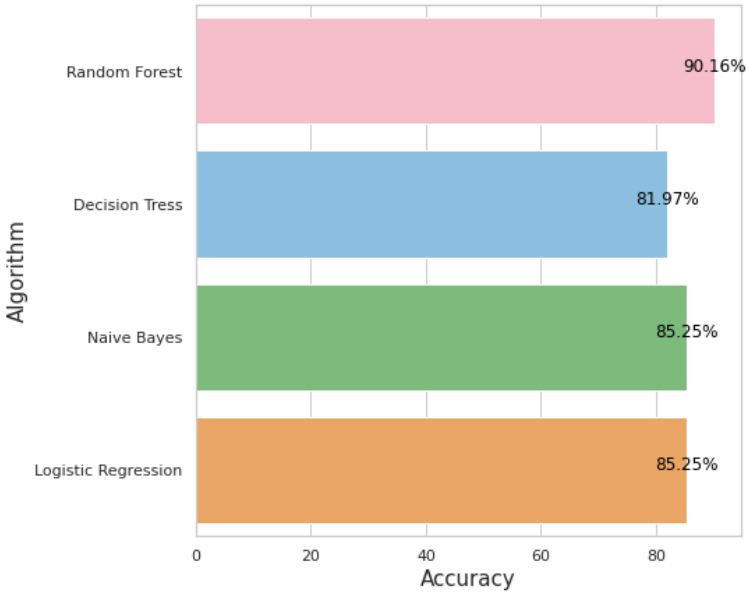
Random Forest > Naive Bayes > Logistic Regression > Decision Tree

همچنین مدل پیاده سازی شده در مقایسه با مقاله (شکل ۱ و ۲) نتایج بهتری داشته است.(Precision, Recall, F1 Score)

۵ جمع بندی

با افزایش تعداد مرگ و میر ناشی از بیماری‌های قلبی ، توسعه سیستمی برای پیش‌بینی موثر و دقیق بیماری‌های قلبی الزامی شده‌است.انگیزه این مطالعه یافتن بهترین الگوریتم ml برای تشخیص بیماری‌های قلبی بود. در این تحقیق با استفاده از داده‌های مخزن یادگیری ماشین UCI ، دقت درخت تصمیم ، رگرسیون لجستیک ، جنگل تصادفی و بیز ساده برای پیش‌بینی بیماری‌های قلبی مقایسه شده‌است. نتایج حاصل از این مطالعه نشان می‌دهد که الگوریتم جنگل تصادفی با نمره صحت ۱۶.۹۰ درصد برای پیش‌بینی بیماری قلبی ، کارآمدترین الگوریتم است(شکل۷). در آینده با توسعه یک کاربرد وب مبتنی بر

الگوریتم جنگل تصادفی و استفاده از مجموعه داده‌های بزرگ‌تر در مقایسه با مجموعه داده‌های مورد استفاده در این تحلیل که به ارائه نتایج بهتر و کمک به متخصصان سلامت در پیش‌بینی بیماری قلبی کمک خواهد کرد ، کار بهبود می‌یابد.



شکل ۷: مقایسه دقت مدل ها