بنام خدا

مقایسه عملکرد سه طبقه بندی کننده برای طبقه بندی مجموعه داده های سرطان سینه

گزارش نهایی و جمع بندی پروژه یادگیری ماشین   
دانشکده علوم ریاضی و کامپیوتر – دانشگاه خوارزمی

تلاشی از سیدسجاد علوی پور

زمستان ۱۴۰۱

هدف : در این گزارش بر اساس بالاترین دقت، ما بهترین مدل را از مدل‌های توصیف‌ شده در این مقاله انتخاب خواهیم کردیم و آن را برای استفاده در آینده برای مجموعه داده بیماران آینده (داده‌های بالینی) استفاده خواهیم نمود.

مقدمه : سرطان سینه یکی از مسائل تهدید کننده زندگی زنان در جامعه و عصر حاضر است. سرطان سینه شایع ترین سرطان در بین زنان است که سالانه ۲.۱ میلیون زن را تحت تاثیر قرار می دهد و همچنین بیشترین مرگ و میر ناشی از سرطان را در بین زنان ایجاد می کند.

در سال ۲۰۱۸ ، تخمین زده می شود که ۶۲۸۰۰۰ زن بر اثر سرطان سینه جان خود را از دست داده اند که تقریباً ۱۵ درصد از کل مرگ و میر ناشی از سرطان در میان زنان است. آمارها نشان می دهد که از ۸۰ درصد یا بیشتر در آمریکای شمالی، سوئد و ژاپن تا حدود ۶۰ درصد در کشورهای با درآمد متوسط و کمتر از ۴۰ درصد در کشورهای کم درآمد متغیر است.

در جامعه ما به دلیل پیروی از فرهنگ های مدرن/غربی و بی توجهی به غذا و عادات زندگی امار مبتلایان به سرطان سینه در حال افزایش است.

بیماری سرطان دارای برخی از سندرم ها را است و بر اساس مطالعه و بررسی این سندرم ها می توانیم تشخیص دهیم که آیا بیمار سرطان سینه دارد یا خیر و با تشخیص خوش‌خیم یا بدخیم بودن سلول سرطانی گامی مثبت برای تشخیص و درمان مبتلایان خواهیم داشت. همچنین این بیماری دارای برخی از سندرم‌ها مانند ضخامت توده، یکنواختی اندازه سلول، یکنواختی شکل سلول، چسبندگی حاشیه‌ای، اندازه سلول اپیتلیال منفرد، هسته‌های برهنه، کروماتین ملایم و هسته‌های نرمال و غیره است.

برای افزایش دقت پیش‌بینی و تشخیص این بیماری، ما می‌توانیم مجموعه داده عظیمی از سرطان سینه داشته باشیم و با استفاده از این داده‌ها می‌توانیم مدل بهتری ایجاد کنیم که بتواند نتیجه یک مجموعه داده ناشناخته را با استفاده از برخی الگوریتم‌های یادگیری ماشین تشخیص دهد.

این به ما کمک می کند تا در زمان سریع یک بیمار را شناسایی کنیم و بتوانیم اقدامات مناسبی را برای این بیمار انجام دهیم.

الگوریتم‌های ماشین بردار پشتیبان ( SVM) ، شبکه عصبی مصنوعی ( ANN) و الگوریتم‌های Naïve Bayes الگوریتم‌های یادگیری نظارت‌شده بسیار محبوب و قدرتمندی برای طبقه‌بندی لیبل و نتیجه ناشناخته هستند.

ما یک مجموعه داده از WBCD (تشخیص سرطان سینه ویسکانسین) را انتخاب کرده‌ایم که شامل ۹ ستون ویژگی و ۱ ستون کلاس و لیبل خروجی است.

ستون های صفت علت و ستون کلاس نتیجه ستون های ویژگی است.

در این مقاله بخش‌های مختلف ماشین بردار پشتیبان (SVM) و الگوریتم (Naïve Bayes ) اساس یک مجموعه داده آموزشی خاص گفته شده آموزش داده‌ایم . همچنین در پردازش اطلاعات و کار با دیتای مورد نظر برای این پرو‌ژه به جهت بالا بردن میزان دقت و صحت و همچنین بررسی متدهای مختلف برای پیش‌بینی و دسته‌بندی داده ها و اطلاعات از روش های دسته بندی و SVM و Naïve Bayes و XGBOOST و Logistic Regression و KNN استفاده کرده‌ایم که ما را به دقت و صحت های به مراتب بهتری نسبت به متدهای پیاده سازی شده مقاله نوشته شده رساند .

رویکرد پیشنهادی :

در روش انجام شده ما به کاوش دیتاست و داده موجود برای بررسی وجود یا عدم وجود بیماری پرداختیم.

برای انجام این امر یک عملیات تحت عنوان پیش‌پردازش (Preprocessing) صورت گرفت و بعد از انجام این عملیات روی مجموعه داده ها شروع به پیاده‌سازی الگوریتم های مورد نظر کردیم .

پیش پردازش : به منظور ارزیابی مدل پیشنهادی، پیش پردازش داده ها به منظور دستیابی به خروجی بهتر با استفاده از مجموعه داده سرطان سینه انجام می شود. یک سری مراحل پیش پردازش انجام می شود تا داده ها در الگوریتم قرار بگیرند و خروجی بهتری داشته باشند.

مراحل به شرح زیر است:

1) اولین ستون ویژگی (شماره شناسه) را حذف می‌کنیم. زیرا به صورت تصادفی تولید می شود و تاثیری بر خروجی ندارد. همچنین ارتباطی با بیماری سرطان سینه ندارد.

۲) برخی از ویژگی‌های گمشده در مجموعه داده وجود دارد و محدوده هر ویژگی 1 تا 10 است. برای این کار، تمام ویژگی‌های از دست رفته را با مقدار یک پر می‌کنیم.

۳) در صفت کلاس، مقدار شامل ۲ و ۴ است. ۲ به معنی خوش خیم (غیر سرطانی) و ۴ به معنای بدخیم (سرطانی) است.

برای سادگی محاسبه، مقدار ۲ را به صفر و مقدار ۴ را به ۱ جایگزین می کنیم.

* در گزارش صورت گرفته نیز قبل از استفاده از دیتاست داده شده پیش پردازش به صورت گفته شده صورت گرفت و داده ها نیز به صورت نمودارهای مختلف با وابستگی به داده کلاس ( نشان دهنده سرطانی و غیرسرطانی ) به خوبی و با دقت بالا به نمایش درآمده‌اند.
* حال به بررسی و مقایسه دقت و صحت الگوریتم های اجرا شده در مقاله و الگوریتم های اجرا شده در گزارش صورت گرفته می‌پردازیم.

۱. الگوریتم اولیه اجرا شده در مقاله و گزارش Classification می‌باشد . که طبق مطلب گفته شده در مقاله ۱۰ درصد از دیتا به عنوان دیتای تست و ۹۰ درصد به عنوان دیتای اموزش و ترین استفاده می‌شوند.

\*درصد دقت گزارش : ۹۲.۸۵ درصد (نکته اصلی و قابل توجه این بود که در صورتی که ۲۵ درصد داده را به تست اختصاص میدادیم میزان دقت و صحت به ۹۶ درصد افزایش پیدا می‌کرد)

۲. دومین الگوریتم انجام شده الگوریتم SVM می‌باشد.

SVM – Kernel = Linear

\*درصد دقت گزارش : ۹۶ درصد

\*درصد دقت مقاله : ۹۶.۷۲ درصد

۳. سومین الگوریتم انجام شده الگوریتم Naïve Bayes می‌باشد.

Gaussian Naïve Bayes

\*درصد دقت گزارش : ۹۵.۴۲۸۵ درصد

\*درصد دقت مقاله : ۹۵.۸۶ درصد

۴. الگوریتم چهارم XGBOOST :

\*درصد دقت گزارش : ۹۷.۸۶ درصد

\*درصد دقت مقاله : ۹۵.۸۸ درصد

۵. الگوریتم پنجم – آزمایشی برای بررسی دقت روش های دیگر :

Logistic Regression

\*درصد دقت گزارش : ۹۶.۵۷ درصد

۶. الگوریتم نهایی - آزمایشی برای بررسی دقت روش های دیگر :

### K-Nearest Neighbors (K-NN)

\*درصد دقت گزارش : ۹۷.۷۱ درصد

نتیجه‌ گیری :

ضرب المثلی هست که پیشگیری بهتر از درمان است. این بدان معناست که بهتر است سعی کنید یک اتفاق بد را از وقوع آن دور نگه دارید تا اینکه یک اتفاق بد را پس از وقوع اصلاح کنید. به همین دلیل بهتر است این مشکل قبل از اینکه به طور جدی در بدن خانم ها اتفاق بیفتد برطرف شود. بنابراین با استفاده از سندرم ها می توان به راحتی بدون مشورت با پزشک پیش بینی کرد که آیا یک زن در وضعیت بهتری قرار دارد یا خیر. از سوی دیگر، به دلیل افزایش تدریجی این بیماری ها، به پزشکان نیز کمک می کند تا به راحتی تعداد زیادی از بیماران را اداره کنند. ما همچنین می توانیم از این برنامه برای دریافت نتایج آماری در مناطق آسیب دیده جهان استفاده کنیم و اقدامات لازم را برای رهایی از سرطان سینه انجام دهیم. با توجه به بسیاری از مسائل اخلاقی در مراقبت های بهداشتی، ما باید بهترین دقت را ارائه دهیم تا در اسرع وقت بهترین گام ها را برای بیماران آسیب دیده برداریم. اما با توجه به ویژگی های داده، انواع مختلف الگوریتم نتایج متفاوتی را برای یک مجموعه داده (WBCD) نشان می دهد. برخی از آنها بسیار خوب هستند و برخی از آنها در نتیجه دارای عملکرد ضعیفی هستند.

در پروژه انجام شده و پیاده سازی الگوریتم های مختلف توانستیم به دقت قابل قبول برای دیتای داده شده برسیم که در برخی موارد از دیتای ابتدایی مقاله نیز بهتر عمل کردند.

در مقاله نوشته شده بالاترین میزان دقت مربوط به Linear SVM می‌باشد با دقت ۹۶.۷۲ که در گزارش نیز به درصد قابل قبول ۹۶ رسیدیم.

ولی در گزارش ما به دقت های بهتری هم رسیدیم ؛ مانند XGBOOST با دقت ۹۷.۸۶ و نزدیک ترین همسایگی با دقت ۹۷.۷۱.