گزارش پروژه پایان فصل درس کاربردهای یادگیری ماشین

عنوان پروژه: تشخیص بیماری پارکینسون در مراحل اولیه با استفاده از الگوریتم‌های یادگیری ماشین به زبان پایتون

نجمه مؤذن ۹۶۱۳۰۹

پاییز ۱۴۰۰

مقدمه:

بیماری پارکینسون (PD) یک اختلال مزمن، طولانی مدت و پیشرونده است که سیستم عصبی مرکزی را تحت تاثیر قرار می‌دهد. علایم بیماری از هر فردی به فرد دیگر کاملا متفاوت است. اولین بار James Parkinson پزشک انگلیسی این بیماری را در سال ۱۸۱۷ توصیف نمود لذا بیماری با نام او شناخته شد.

نشانه‌های این بیماری معمولاً آرام و به‌تدریج ظاهر می‌شوند و با پیشرفت بیماری، علائم غیرحرکتی نیز بروز می‌کنند. آشکارترین نشانه‌های زودرس این بیماری عبارتند از لرزش، خشکی بدن، آرام‌شدن حرکات، و دشواری در راه رفتن. نشانه‌های شناختی و رفتاری این بیماری نیز در اغلب افراد معمولاً به شکل افسردگی، اضطراب و فقدان علاقه و هیجان بروز می‌کند. در مراحل پیشرفته بیماری پارکینسون، بعضاً زوال عقل نیز شایع است. فرد مبتلا به پارکینسون ممکن است مشکلاتی در خوابیدن و سیستم حواس خود نیز تجربه کند.

نشانه‌های حرکتی این بیماری به علت از بین رفتن سلول‌ها در توده سیاه مغز و در نتیجه آن کاهش دوپامین (که یک انتقال دهنده عصبی است) رخ می‌دهد. دوپامین برای حفظ الگوهای حرکتی طبیعی بدن اهمیت زیادی دارد. دقیقاً به همین دلیل است که بسیاری از درمان‌های پارکینسون با هدف افزایش سطح دوپامین در مغز انجام می‌شوند.

اختلال در ماده سیاه و اختلال در تولید دوپامین سلول های مغزی، منجر به سه علامت اصلی در پارکینسون می گردد:

- کند شدن حرکت: برای مثال، راه رفتن و یا بلند شدن از روی یک صندلی می توانند دشوار تر شوند. هنگامی که این علامت برای اولین بار ظاهر می شود، ممکن است فرد به اشتباه فکر کند که فقط "پیر شده است". تشخیص بیماری پارکینسون بدون ظهور علائم دیگر آن ممکن نیست. با مرور زمان، فرد دچار یک الگوی حرکتی می شود که در آن با حالت موج گونه و بی قرار راه رفته و شروع، پایان و یا تغییر جهت در طول راه رفتن برای او دشوار خواهد شد.

- خشکی و سفتی اندامها (سفت شدن عضلات دستها و پاها): عضلات فرد ممکن است بی قرار شوند. همچنین به هنگام راه رفتن، دست های شما به مقدار طبیعی جلو و عقب نخواهند رفت.

- لرزیدن (رعشه): این علامت بسیار شایع است ولی در همه ی موارد رخ نخواهد داد. این عامل معمولا انگشتان و کل دست را تحت تاثیر قرار می دهد ولی می تواند در دیگر نقاط بدن نیز ظاهر شود و به هنگام استراحت فرد، راحت تر می توان آن را تشخیص داد. این علامت بیماری می تواند به هنگام احساساتی یا مضطرب شدن فرد، تشدید شده و به هنگام انجام برخی از کار ها مانند برداشتن یک جسم، کاهش یابد.
- اختلال در راه رفتن و تعادل وضعیتی: عادل وضعیی ممکن است در پارکینسون باشد یا نباشد و جزئ علایم اصلی نیست.

اگر برنامه ای برای کنترل این علائم وجود نداشته باشد با گذشت زمان و به تدریج علایم بدتر خواهند شد.

در بیماری پارکینسون، از بین رفتن نورون‌ها در سایر بخش‌های مغز هم رخ می‌دهد، و زمینه‌ساز برخی علائم غیرحرکتی این بیماری می‌شود. اگرفردی به بیماری پارکینسون مبتلا شود، تعدادی از سلول های توده ی سیاه آسیب دیده و می میرند. علت دقیق این امر هنوز به درستی معلوم نیست. با گذشت زمان، تعداد بیشتری از این سلول ها آسیب دیده و خواهند مرد. با آسیب دیدن این سلول ها، مقدار دوپامین تولید شده کاهش می یابد. کاهش تعداد سلول ها به همراه کاهش میزان دوپامین در سلول های این بخش از مغز، می تواند باعث کند و غیر طبیعی شدن پیام های عصبی فرستاده شده به سمت ماهیچه ها گردد.

علائم دیگری نیز ممکن است در اثر مشکلاتی که سلول ها ی مغزی و اعصاب با کنترل عضلات دارند، ایجاد شوند. این علائم عبارت اند از:

- کاهش حالات در چهره از جمله لبخند یا اخم. کاهش پلک زدن.
- دشواری در انجام حرکات ظریف مانند بستن بند کفش و یا بستن دکمه های پیراهن.
- دشواری در نوشتن با دست (دست خط فرد ریز تر خواهد شد).
- دشواری در ایجاد تعادل و حفظ حالت مناسب بدن و همچنین افزایش احتمال زمین خوردن.
- حرف زدن فرد ممکن است کند و خسته کننده و یکنواخت شود.
- صدای فرد دارای لرزش و خش دار می‌شود.
- بلعیدن می تواند برای فرد دشوار شده و بزاق در دهان تجمع یابد.
- خستگی و احساس درد
- درد کف پا

تعداد زیادی از علائم دیگر نیز در برخی از موارد و معمولا با تشدید بیماری به وجود می آیند. این علائم شامل موارد زیر می باشند:

- یبوست.
- مشکلات مربوط به مثانه و بعضا بی اختیاری.
- توهم (دیدن، شنیدن و یا بوییدن چیز هایی که واقعا وجود ندارند).

- عرق کردن.
- مشکلات مربوط به حس بویایی.
- مشکلات مربوط به خواب.
- کاهش وزن.
- درد.
- افسردگی.
- اضطراب.
- دشواری در کنترل انگیزه ها. برای مثال، خوردن، خریدن کردن.

علاوه بر کاهش دوپامین و سلول‌هایی که دوپامین می‌سازند، پروتئینی به نام آلفا-ساینوکلین هم در بیماری پارکینسون نقش دارند. آلفا-سینوکلئین در حالت عادی به برقراری ارتباط نورون‌ها با یکدیگر کمک می‌کند اما در بیماری پارکینسون، این پروتئین در توده‌های میکروسکوپی به نام جسم لویی کپه‌کپه جمع می‌شود. محققان بر این باورند که آلفا-سینوکلئین در پیشرفت پارکینسون نقش دارد و شاید بتوان درمان‌های جدیدی ایجاد کرد که جلوی جمع‌شدن این پروتئین را بگیرد.

اگرچه علت بروز این بیماری کاملاً مشخص نیست، اما می‌دانیم که هر دو فاکتور ژنتیکی و محیطی در بروز آن نقش دارند. از این میان می‌توان به فاکتورهایی مانند افزایش سن، پیشینه خانوادگی ابتلا به بیماری پارکینسون، جهش‌های ژنتیکی، جنسیت، قرار گرفتن در معرض سموم دفع آفات، برخی داروها و پیشینه آسیب و جراحت مغز به عنوان عوامل بروز این بیماری اشاره کرد.

با توجه به بررسی‌های بالینی که بر روی بیماران مبتلابه پارکینسون انجام‌شده، پیشرفت این بیماری در ۵ مرحله صورت می‌گیرد. این طبقه‌بندی و آشنایی با مراحل، به پزشکان کمک می‌کند تا متوجه شوند که بیمار در چه مرحله‌ای قرار دارد تا درمان‌های لازم را تجویز کنند. مراحل بیماری پارکینسون شامل موارد زیر است:

مرحله اول

مراحل اولیه بیماری پارکینسون به‌صورت عمده شامل بروز و شناسایی اولیه علائم بیماری است. در مرحله اول اغلب نشانه‌های کمی از بیماری مشاهده‌شده که معمولاً در بین بیماران، لرزش یکی از دست‌ها یا پاها بسیار رایج است . در این مرحله از بیماری، ممکن است عدم تعادل خفیف درراه رفتن یا ایستادن بیمار مشاهده شود، یا در برخی مواقع حالات چهره و صورت او تغییر کند. البته که این نشانه‌ها معمولاً موجب اختلال در زندگی روزانه فرد نمی‌شوند؛ اما بهتر است که توسط پزشک متخصص مغز و اعصاب تشخیص داده و درمان آغاز شود.

مرحله دوم

در مرحله دوم از مراحل بیماری پارکینسون، علائم بیماری شدت بیشتری پیدا می‌کند. ورود به این مرحله متناسب با شرایط جسمی و روحی بیمار ممکن است چند ماه یا چند سال طول بکشد؛ بنابراین، توجه به وضعیت بیمار در مرحله نخست و این مرحله بسیار لازم و حیاتی است. افزایش شدت نشانه‌ها به‌صورت سفت شدن عضلات دو طرف بدن در قسمت دست‌ها یا پاها، ایجاد مشکل در تکلم و

حفظ تعادل بدن هنگام ایستادن، به‌عنوان بعضی از علائم مرحله دوم پارکینسون به شمار می‌روند. در این مرحله هم با کنترل بیماری، فرد می‌تواند به انجام کارهای روزانه خود ادامه دهد.

مرحله سوم

این مرحله از مراحل بیماری پارکینسون، مرحله میانی رشد پارکینسون به شمار می‌رود. در این مرحله شدت علائم مرحله دوم بیشتر می‌شود. همچنین نشانه‌های دیگری مثل عدم تعادل هنگام انجام کارهای روزانه، واکنش‌های کند و آرام نسبت به اتفاقات محیطی نیز دیده‌شده و به‌طورکلی حرکات بیمار بسیار کند می‌شود. علاوه بر دارودرمانی از فیزیوتراپی نیز در این مرحله استفاده می‌شود. در مرحله سوم پارکینسون، بیمار نیاز به کمک و توجه بیشتری برای انجام کارهای شخصی مانند لباس پوشیدن، راه رفتن و غذا پختن دارد.

مرحله چهارم

مرحله چهارم و پنجم از مراحل آخر بیماری پارکینسون به‌حساب می‌آیند. در مرحله چهارم، خشکی عضلات و کند شدن حرکات بدنی افزایش‌یافته و در مواردی ممکن است شدت لرزش بدن کاهش یابد؛ اما بااین‌وجود راه رفتن و حفظ تعادل بدن بسیار سخت است .در این مرحله از بیماری، بهتر است همواره فردی در کنار بیمار باشد تا در انجام کارهای روزانه به او کمک کند. همچنین ادامه داشتن جلسات فیزیوتراپی و استفاده از واکر هنگام راه رفتن می‌تواند مفید باشد.

مرحله پنجم

مرحله پنجم که آخرین مرحله از پیشرفت بیماری پارکینسون است، شامل علائم مزمن بیماری است. در این مرحله سفت شدن عضلات و ناتوانی درحرکت دیده می‌شود. مراقبت روزانه در مرحله آخر بسیار واجب و حیاتی است؛ زیرا اغلب بیماران به دلیل مشکلات حرکتی قادر به انجام فعالیت‌های روزانه خود نیستند .استفاده از ویلچر، واکر و فیزیوتراپی بسیار کمک‌کننده است. برخی مشکلات روان‌شناختی و عصبی مثل توهم، گیجی، مشکل در بلع مواد غذایی و زوال عقل نیز ممکن است در بیماران مشاهده شود.

اهمیت موضوع:

بیماری پارکینسون پس از آلزایمر شایع‌ترین بیماری مخرب اعصاب به شمار می‌رود. پارکینسون معمولاً در افراد بالای ۶۰ سال بروز می‌کند. مردان بیشتر از زنان به این بیماری مبتلا می‌شوند و نسبت ابتلای مردان به زنان ۳ به ۲ است. پارکینسون ممکن است در افراد زیر ۵۰ سال نیز ایجاد شود که در این صورت پارکینسون زودرس خوانده می‌شود. طبق آمار سال ۲۰۱۵، ۲/۶ میلیون فرد در جهان مبتلا به پارکینسون هستند و این بیماری سالانه به ۱۱۷۴۰۰ مرگ منجر می‌شود.

به صورت دائمی برای بیماری پارکینسون درمانی وجود ندارد و هیچگونه روشی نمیتواند در جلوگیری از پیشرفت آن تاثیری داشته باشد اما روش‌های مداوا برای کاهش آثار این بیماری وجود دارند. مداوای اولیه معمولاً با تجویز دارو شروع می‌شود. همچنین فیزیوتراپی، کاردرمانی و ورزش درمانی، تحریک مغناطیسی مغز، استفاده از امواج اولتراساند به کمک دستگاه MRI و در موارد پیشرفته‌تر انجام عمل جراحی روش‌هایی هستند که امروزه برای کنترل بیماری پارکینسون استفاده می‌شوند. مشخص شده‌است که رژیم غذایی در بهبود این علائم تا حدی مؤثر است.

بنابراین تشخیص این بیماری در مراحل اولیه بسیار حائز اهمیت است. چرا که بیمار علائم اولیه و نامحسوس را به سن زیاد نسبت می‌دهد. معمولا بیمار وقتی به پزشک مراجعه می‌کند که بیماری به مراحل پیشرفته خود رسیده و کنترل آن بسیار مشکل و بعضا غیرممکن خواهد بود.

از طرفی هیچ آزمایش خاصی برای تشخیص بیماری پارکینسون وجود ندارد. پزشکی که در زمینه بیماری های سیستم عصبی (متخصص مغز و اعصاب) آموزش دیده است، بیماری پارکینسون را بر اساس تاریخچه پزشکی، بررسی علائم و نشانه های بیمار و معاینه بالینی و عصبی تشخیص می دهد. باید توجه داشت که علائم اولیه بیماری حتی برای پزشک نیز به سختی قابل تشخیص نیست و گاهی اوقات برای تشخیص بیماری پارکینسون توسط پزشک زمان لازم است تا ارزیابی وضعیت و علائم بیمار در طول زمان مشخص شود و این به ضرر بیمار است. پزشک ممکن است اسکن SPECT (توموگرافی کامپیوتری تک فوتونی) را که به آن اسکن انتقال دهنده دوپامین (DAT) می گویند، پیشنهاد کند. اگرچه این می تواند به تایید تشخیص کمک کند، اما علائم و معاینه عصبی در نهایت تشخیص صحیح را تعیین می کنند. بیشتر افراد به اسکن DAT احتیاج ندارند. روش های تصویربرداری مانند MRI ، CT، سونوگرافی مغز و PETاسکن همچنین ممکن است برای کمک به رد سایر اختلالات استفاده شود. روش های تصویربرداری به طور اختصاصی برای تشخیص بیماری پارکینسون مفید نیستند.

در مجموع همه این‌ها به این معناست که ما در تشخیص بیماری پارکینسون در بیماران، دیر عمل می کنیم.

هدف پروژه:

تشخیص به موقع بیماری پارکینسون، با تمام پیشرفت‌های علمی، هنوز هم به عنوان یک چالش باقی مانده است. در سال‌های اخیر، محققان برای تشخیص این بیماری در مراحل اولیه به یادگیری ماشین و الگوریتم‌های مربوط به طبقه‌بندی روی آورده‌اند. یعنی با استفاده از الگوریتم‌ها و اعمال آن‌ها به مجموعه داده‌های مختلف می‌توان تشخیص داد که آیا فرد مشکوک به پارکینسون، بیمار است یا خیر.

هدف از این پروژه ارائه روش‌هایی ساده و کم هزینه با دقت بالا برای تشخیص زودهنگام بیماری پارکینسون است.

مجموعه داده‌ها:

برای طراحی یک سیستم تشخیص‌گر دقیق برای تشخیص بیماری پارکینسون به دیتاست‌هایی نیاز است که شامل علائم این بیماری که در بخش‌های قبل ذکر شد، باشد. به این منظور در این بخش به چند دسته از دیتاست‌های موجود برای بیماری پارکینسون اشاره خواهیم کرد:

- Gait dataset: این دیتاست مربوط به داده‌هایی است که از نحوه راه رفتن افراد گرفته می‌شود. به این منظور سنسورهایی را به بدن بیمار وصل می‌کنند، و از شخص می‌خواهند که راه برود و اینچنین داده‌گیری می‌کنند. این پایگاه داده شامل معیارهای

راه رفتن از ۹۳ بیمار مبتلا به PD (میانگین سن: ۶۶,۳ سال؛ ۶۳٪ مرد) و ۷۳ فرد سالم (میانگین سن: ۶۶,۳ سال؛ ۵۵٪ مرد) است. پایگاه داده شامل سوابق نیروی واکنش عمودی زمین از سوژه‌ها است که با سرعت معمول و انتخابی خود برای تقریباً ۲ دقیقه روی زمین هموار راه می‌رفتند. در زیر هر پا ۸ حسگر وجود داشت که نیرو (بر حسب نیوتن) را بر حسب زمان اندازه گیری می کردند. خروجی هر یک از این ۱۶ حسگر دیجیتالی و با ۱۰۰ نمونه در ثانیه ثبت شده است و همچنین رکوردها شامل دو سیگنال است که مجموع ۸ خروجی سنسور را برای هر پا منعکس می کند.

- Spiral drawings dataset: این دیتاست شامل داده‌هایی تصویری از نقاشی‌های مارپیچ و موجی افراد است. پایگاه داده کنترل دستخط PD شامل ۶۲ فرد PWP (افراد مبتلا به پارکینسون) و ۱۵ فرد سالم است. از همه افراد، سه نوع ضبط دست خط آزمون مارپیچ ایستا (SST)، تست مارپیچی پویا (DST) و تست پایداری در نقطه معین (STCP) گرفته می شود.
- Voice dataset: این دیتاست شامل داده‌های صوتی از افراد است. حدود شش ضبط برای هر بیمار وجود دارد، نام بیمار در ستون اول مشخص شده است. در این مجموعه داده هر ستون در جدول یک معیار صوتی خاص است و هر ردیف مربوط به یکی از ۱۹۵ صدای ضبط شده از این افراد است (ستون "name"). این دیتاست شامل ۱۹۵ سیگنال (۱۴۷ بیمار و ۴۸ سالم) است که از هر سیگنال ۲۲ ویژگی استخراج شده است. هدف اصلی داده‌ها این است که افراد سالم را از افراد مبتلا به PD بر اساس ستون "status" که ۰ برای سالم و ۱ برای PD تنظیم شده است، متمایز کند.

در این پروژه از voice dataset استفاده شده است.

الگوریتم‌های یادگیری ماشین:

از بین ۱۹۵ کیسی که در ویس دیتاست مذکور وجود دارد، ۱۴۷ مورد بیمارند و ۴۸ مورد سالم هستند. ما ۸۰ درصد از دیتاها را برای train قرار می‌دهیم و بقیه را برای test. این یعنی که ۱۵۶ مورد برای train وجود دارد و ۳۹ مورد برای test. توجه کنیم که همانطور که گفته شد، کلاس ۰ برای سالم و ۱ برای PD تنظیم شده است.

در این پروژه برای تشخیص بیماری پارکینسون برای ویس دیتاستی که در بالا معرفی شد، ۹ الگوریتم‌های طبقه‌بندی به کار گرفته شده است که عبارتند از:

1. Linear Regression
2. Logistic Regression
3. Support Vector Machine
4. Decision Tree
5. Random Forrest
6. Extreme Gradient Boosting
7. K-Nearest Neighbors

8. Naïve Bayes
9. Neural Network

تفسیر نتایج:

در این پروژه الگوریتم‌های مذکور پیاده شدند و سپس برای هرکدام از این الگوریتم‌ها accuracy، recall و precision محاسبه شد و همچنین confusion matrix نیز بدست آورده شد و heat map مربوط به آن نیز رسم شده است.

حال به طور جزئی به بررسی نتایج هریک از الگوریتم‌ها می‌پردازیم:

- **Linear Regression:**

با رسم confusion matrix heat map برای این الگوریتم Fig1 را خواهیم داشت.

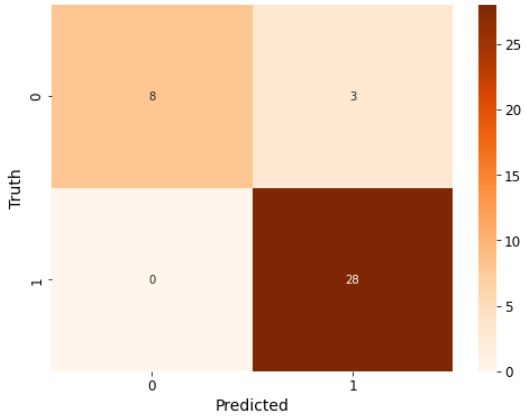


Fig1: Linear Regression heat map

این heat map، ماتریس درهم ریختگی یا confusion matrix را نشان می‌دهد که می‌گوید چه تعداد از کل پیش‌بینی‌ها درست بوده است. محور افقی پیش‌بینی‌ها را نشان می‌دهد و محور عمودی واقعیت را. درایه‌های روی قطر تعداد پیش‌بینی‌های درست را نشان می‌دهد. یعنی یعنی ازمیان ۳۹ مورد تست شده، ۸ نفر واقعا سالم هستند و ماشین هم به درستی پیش‌بینی کرده که سالم‌اند. ۲۸ نفر واقعا بیمارند و ماشین هم به درستی پیش‌بینی کرده که بیمارند. از طرفی ماشین هیچ پیش‌بینی اشتباهی در این مورد که فرد واقعا بیمار بوده و ماشین سالم پیش‌بینی کند، نداشته است. همچنین ۳ نفر واقعا سالم‌اند اما ماشین پیش‌بینی کرده که بیمارند.

خروجی دیگر بدست آمده برای این الگوریتم چنین است:

```
confusion_matrix
[[ 8  3]
 [ 0 28]]
```

accuracy: 0.9230769230769231

recall: [0.72727273 1.]

precision: [1. 0.90322581]

مقدار accuracy درصد درستی و دقت پیش‌بینی‌ها را نشان می‌دهد. برای این روش این مقدار تقریبا ۰/۹۲ بدست آمده است و این یعنی که حدود ۹۲ درصد پیش بینی‌های انجام شده درست هستند و ۸ درصد نادرستند.

مقدار recall نشان می‌دهد که چند درصد از داده‌های مربوط به هر کلاس را به درستی توانستیم پیش‌بینی کنیم. یعنی در این روش برای کلاس ۰ (سالم)، حدود ۷۲ درصد و برای کلاس ۱ (بیمار)، ۱۰۰ درصد از پیش‌بینی‌هایمان درستند.

مقدار precision نشان می‌دهد که چند درصد از از پیش‌بینی‌های انجام شده برای هرکلاس، واقعا در آن کلاس بوده اند. یعنی برای کلاس ۰ (سالم)، ۱۰۰ درصد از پیش‌بینی‌هایی که انجام شده واقعا در کلاس ۰ (سالم) بوده اند. همچنین برای کلاس ۱ (بیمار)، ۹۰ درصد از پیش‌بینی‌هایی که انجام شده واقعا در کلاس ۱ (بیمار) بوده اند.

- **Logistic Regression:**

با رسم confusion matrix heat map برای این الگوریتم Fig2 را خواهیم داشت.

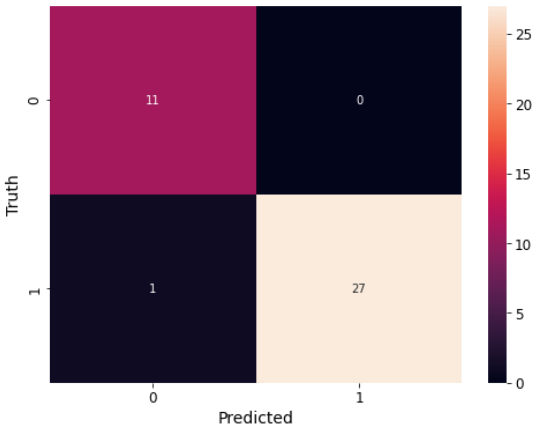


Fig2: Logistic Regression heat map

این heat map، ماتریس درهم ریختگی یا confusion matrix را نشان می‌دهد که می‌گوید چه تعداد از کل پیش‌بینی‌ها درست بوده است. محور افقی پیش‌بینی‌ها را نشان می‌دهد و محور عمودی واقعیت را. درایه‌های روی قطر تعداد پیش‌بینی‌های درست را نشان می‌دهد. یعنی ازمیان ۳۹ مورد تست شده، ۱۱ نفر واقعا سالم هستند و ماشین هم به درستی پیش‌بینی کرده که سالم‌اند. ۲۷ نفر واقعا بیمارند و ماشین هم به درستی پیش‌بینی کرده که بیمارند. از طرفی ۱ نفر واقعا بیمار بوده و ماشین اشتباها آن را سالم پیش‌بینی کرده است. همچنین ۰ نفر واقعا سالم‌اند اما ماشین پیش‌بینی کرده که بیمارند.

خروجی دیگر بدست آمده برای این الگوریتم چنین است:

confusion_matrix

[[11 0]
 [1 27]]

accuracy: 0.9743589743589743

recall: [1.   0.96428571]

precision: [0.91666667 1.   ]

مقدار accuracy درصد درستی و دقت پیش‌بینی‌ها را نشان می‌دهد. برای این روش این مقدار تقریبا ۰/۹۷ بدست آمده است و این یعنی که حدود ۹۷ درصد پیش بینی‌های انجام شده درست هستند و تنها ۳ درصد نادرستند.

مقدار recall نشان می‌دهد که چند درصد از داده‌های مربوط به هر کلاس را به درستی توانستیم پیش‌بینی کنیم. یعنی در این روش برای کلاس ۰ (سالم)، ۱۰۰ درصد و برای کلاس ۱ (بیمار)، ۹۶ درصد از پیش‌بینی‌هایمان درستند.

مقدار precision نشان می‌دهد که چند درصد از از پیش‌بینی‌های انجام شده برای هرکلاس، واقعا در آن کلاس بوده اند. یعنی برای کلاس ۰ (سالم)، ۹۱ درصد از پیش‌بینی‌هایی که انجام شده واقعا در کلاس ۰ (سالم) بوده‌اند. همچنین برای کلاس ۱ (بیمار)، ۱۰۰ درصد از پیش‌بینی‌هایی که انجام شده واقعا در کلاس ۱ (بیمار) بوده اند.

- **Decision Tree:**

با رسم confusion matrix heat map برای این الگوریتم Fig3 را خواهیم داشت.

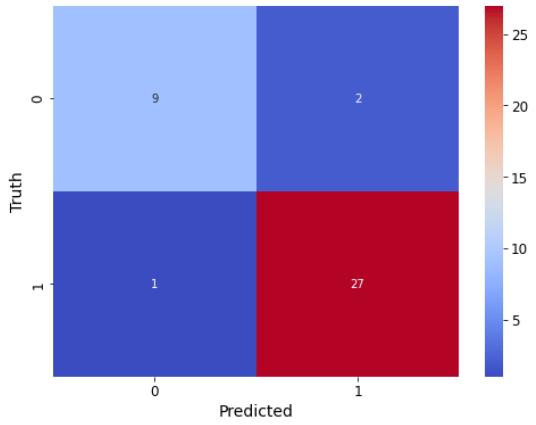


Fig3: Decision Tree heat map

این heat map، ماتریس درهم ریختگی یا confusion matrix را نشان می‌دهد که می‌گوید چه تعداد از کل پیش‌بینی‌ها درست بوده است. محور افقی پیش‌بینی‌ها را نشان می‌دهد و محور عمودی واقعیت را. درایه‌های روی قطر تعداد پیش‌بینی‌های درست را نشان می‌دهد. یعنی یعنی ازمیان ۳۹ مورد تست شده، ۹ نفر واقعا سالم هستند و ماشین هم به درستی پیش‌بینی کرده که سالم‌اند. ۲۷ نفر واقعا بیمارند و ماشین هم به درستی پیش‌بینی کرده که بیمارند. از طرفی ۱ نفر واقعا بیمار بوده و ماشین اشتباها آن را سالم پیش‌بینی کرده است. همچنین ۲ نفر واقعا سالم‌اند اما ماشین پیش‌بینی کرده که بیمارند.

خروجی دیگر بدست آمده برای این الگوریتم چنین است:

```
confusion_matrix
[[ 9  2]
 [ 1 27]]

accuracy: 0.9230769230769231

recall: [0.81818182  0.96428571]

precision: [0.9       0.93103448]
```

مقدار accuracy درصد درستی و دقت پیش‌بینی‌ها را نشان می‌دهد. برای این روش این مقدار تقریبا ۰/۹۲ بدست آمده است و این یعنی که حدود ۹۲ درصد پیش بینی‌های انجام شده درست هستند و تنها ۷ درصد نادرستند.

مقدار recall نشان می‌دهد که چند درصد از داده‌های مربوط به هر کلاس را به درستی توانستیم پیش‌بینی کنیم. یعنی در این روش برای کلاس ۰ (سالم)، حدود ۸۱ درصد و برای کلاس ۱ (بیمار)، ۹۶ درصد از پیش‌بینی‌هایمان درستند.

مقدار precision نشان می‌دهد که چند درصد از از پیش‌بینی‌های انجام شده برای هرکلاس، واقعا در آن کلاس بوده اند. یعنی برای کلاس ۰ (سالم)، ۹۰ درصد از پیش‌بینی‌هایی که انجام شده واقعا در کلاس ۰ (سالم) بوده‌اند. همچنین برای کلاس ۱ (بیمار)، حدود ۹۳ درصد از پیش‌بینی‌هایی که انجام شده واقعا در کلاس ۱ (بیمار) بوده اند.

- **Support Vector Machine:**

با رسم confusion matrix heat map برای این الگوریتم Fig4 را خواهیم داشت.

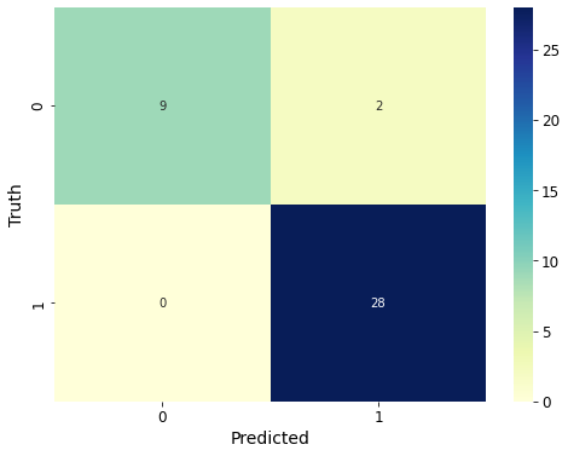


Fig4: Support Vector Machine heat map

این heat map، ماتریس درهم ریختگی یا confusion matrix را نشان می‌دهد که می‌گوید چه تعداد از کل پیش‌بینی‌ها درست بوده است. محور افقی پیش‌بینی‌ها را نشان می‌دهد و محور عمودی واقعیت را. درایه‌های روی قطر تعداد پیش‌بینی‌های

درست را نشان می‌دهد. یعنی یعنی ازمیان ۳۹ مورد تست شده، ۹ نفر واقعا سالم هستند و ماشین هم به درستی پیش‌بینی کرده که سالم‌اند. ۲۸ نفر واقعا بیمارند و ماشین هم به درستی پیش‌بینی کرده که بیمارند. از طرفی ۰ نفر واقعا بیمار بوده و ماشین اشتباها آن را سالم پیش‌بینی کرده است. همچنین ۲ نفر واقعا سالم‌اند اما ماشین پیش‌بینی کرده که بیمارند.

خروجی دیگر بدست آمده برای این الگوریتم چنین است:

```
confusion_matrix
[[ 9  2]
 [ 0 28]]

accuracy: 0.9487179487179487

recall: [0.81818182       1.       ]

precision: [1.         0.93333333]
```

مقدار accuracy درصد درستی و دقت پیش‌بینی‌ها را نشان می‌دهد. برای این روش این مقدار تقریبا ۰/۹۴ بدست آمده است و این یعنی که حدود ۹۴ درصد پیش بینی‌های انجام شده درست هستند و تنها ۶ درصد نادرستند.

مقدار recall نشان می‌دهد که چند درصد از داده‌های مربوط به هر کلاس را به درستی توانستیم پیش‌بینی کنیم. یعنی در این روش برای کلاس ۰ (سالم)، حدود ۸۱ درصد و برای کلاس ۱ (بیمار)، ۱۰۰ درصد از پیش‌بینی‌هایمان درستند.

مقدار precision نشان می‌دهد که چند درصد از از پیش‌بینی‌های انجام شده برای هرکلاس، واقعا در آن کلاس بوده اند. یعنی برای کلاس ۰ (سالم)، ۱۰۰ درصد از پیش‌بینی‌هایی که انجام شده واقعا در کلاس ۰ (سالم) بوده‌اند. همچنین برای کلاس ۱ (بیمار)، حدود ۹۳ درصد از پیش‌بینی‌هایی که انجام شده واقعا در کلاس ۱ (بیمار) بوده اند.

- **Random Forrest**:

با رسم confusion matrix heat map برای این الگوریتم Fig5 را خواهیم داشت.

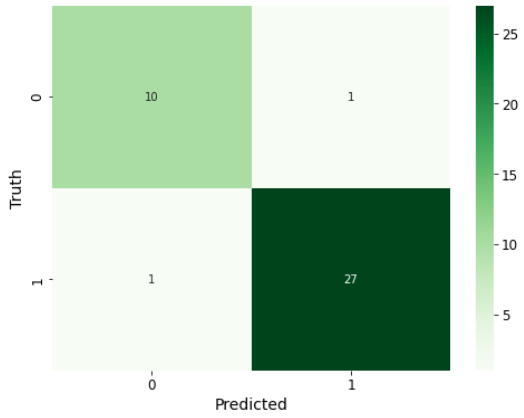


Fig5: Random Forrest heat map

این heat map، ماتریس درهم ریختگی یا confusion matrix را نشان می‌دهد که می‌گوید چه تعداد از کل پیش‌بینی‌ها درست بوده است. محور افقی پیش‌بینی‌ها را نشان می‌دهد و محور عمودی واقعیت را. درایه‌های روی قطر تعداد پیش‌بینی‌های درست را نشان می‌دهد. یعنی یعنی ازمیان ۳۹ مورد تست شده، ۱۰ نفر واقعا سالم هستند و ماشین هم به درستی پیش‌بینی کرده که سالم‌اند. ۲۷ نفر واقعا بیمارند و ماشین هم به درستی پیش‌بینی کرده که بیمارند. از طرفی ۱ نفر واقعا بیمار بوده و ماشین اشتباها آن را سالم پیش‌بینی کرده است. همچنین ۱ نفر واقعا سالم‌اند اما ماشین پیش‌بینی کرده که بیمارند.

خروجی دیگر بدست آمده برای این الگوریتم چنین است:

```
confusion_matrix
[[10  1]
 [ 1 27]]

accuracy: 0.9487179487179487

recall: [0.90909091  0.96428571]

precision: [0.90909091  0.96428571]
```

مقدار accuracy درصد درستی و دقت پیش‌بینی‌ها را نشان می‌دهد. برای این روش این مقدار تقریبا ۰/۹۴ بدست آمده است و این یعنی که حدود ۹۴ درصد پیش بینی‌های انجام شده درست هستند و تنها ۶ درصد نادرستند.

مقدار recall نشان می‌دهد که چند درصد از داده‌های مربوط به هر کلاس را به درستی توانستیم پیش‌بینی کنیم. یعنی در این روش برای کلاس ۰ (سالم)، حدود ۹۰ درصد و برای کلاس ۱ (بیمار)، ۹۶ درصد از پیش‌بینی‌هایمان درستند.

مقدار precision نشان می‌دهد که چند درصد از از پیش‌بینی‌های انجام شده برای هرکلاس، واقعا در آن کلاس بوده اند. یعنی برای کلاس ۰ (سالم)، ۹۰ درصد از پیش‌بینی‌هایی که انجام شده واقعا در کلاس ۰ (سالم) بوده‌اند. همچنین برای کلاس ۱ (بیمار)، حدود ۹۶ درصد از پیش‌بینی‌هایی که انجام شده واقعا در کلاس ۱ (بیمار) بوده اند.

- **Extreme Gradient Boosting:**

با رسم confusion matrix heat map برای این الگوریتم Fig6 را خواهیم داشت.

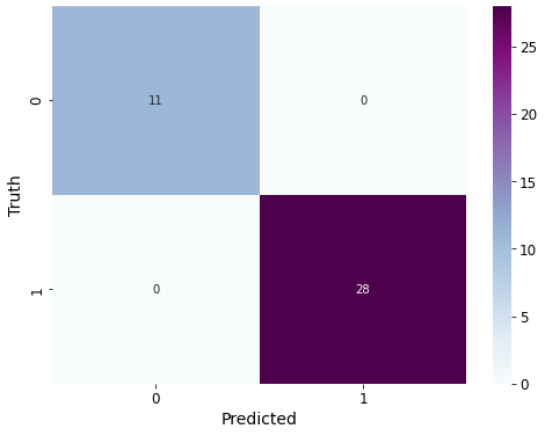


Fig6: Extreme Gradient Boosting heat map

این heat map، ماتریس درهم ریختگی یا confusion matrix را نشان می‌دهد که می‌گوید چه تعداد از کل پیش‌بینی‌ها درست بوده است. محور افقی پیش‌بینی‌ها را نشان می‌دهد و محور عمودی واقعیت را. درایه‌های روی قطر تعداد پیش‌بینی‌های درست را نشان می‌دهد. یعنی یعنی ازمیان ۳۹ مورد تست شده، ۱۱ نفر واقعا سالم هستند و ماشین هم به درستی پیش‌بینی کرده که سالم‌اند. ۲۸ نفر واقعا بیمارند و ماشین هم به درستی پیش‌بینی کرده که بیمارند. از طرفی ۰ نفر واقعا بیمار بوده و ماشین اشتباها آن را سالم پیش‌بینی کرده است. همچنین ۰ نفر واقعا سالم‌اند اما ماشین پیش‌بینی کرده که بیمارند.

خروجی دیگر بدست آمده برای این الگوریتم چنین است:

```
confusion_matrix
[[11  0]
 [ 0 28]]

accuracy: 1.0

recall: [1. 1.]

precision: [1. 1.]
```

مقدار accuracy درصد درستی و دقت پیش‌بینی‌ها را نشان می‌دهد. برای این روش این مقدار ۱ بدست آمده است و این یعنی که ۱۰۰ درصد پیش بینی‌های انجام شده درست هستند و ۰ درصد نادرستند.

مقدار recall نشان می‌دهد که چند درصد از داده‌های مربوط به هر کلاس را به درستی توانستیم پیش‌بینی کنیم. یعنی در این روش برای کلاس ۰ (سالم)، ۱۰۰درصد و برای کلاس ۱ (بیمار)، ۱۰۰ درصد از پیش‌بینی‌هایمان درستند.

مقدار precision نشان می‌دهد که چند درصد از از پیش‌بینی‌های انجام شده برای هرکلاس، واقعا در آن کلاس بوده اند. یعنی برای کلاس ۰ (سالم)، ۱۰۰ درصد از پیش‌بینی‌هایی که انجام شده واقعا در کلاس ۰ (سالم) بوده‌اند. همچنین برای کلاس ۱ (بیمار)، حدود ۱۰۰ درصد از پیش‌بینی‌هایی که انجام شده واقعا در کلاس ۱ (بیمار) بوده اند.

- **Artificial Neural Network**:

با رسم confusion matrix heat map برای این الگوریتم Fig7 را خواهیم داشت.

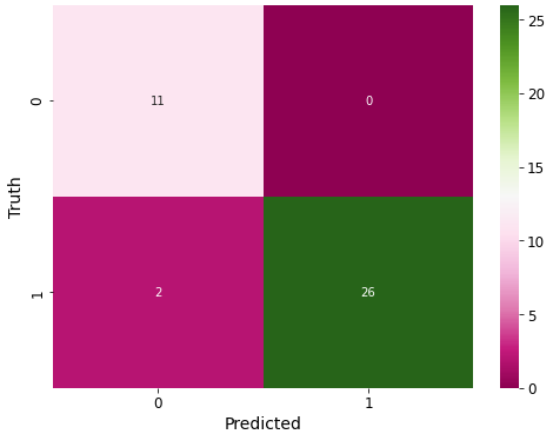


Fig7: Artificial Neural Network heat map

این heat map، ماتریس درهم ریختگی یا confusion matrix را نشان می‌دهد که می‌گوید چه تعداد از کل پیش‌بینی‌ها درست بوده است. محور افقی پیش‌بینی‌ها را نشان می‌دهد و محور عمودی واقعیت را. درایه‌های روی قطر تعداد پیش‌بینی‌های درست را نشان می‌دهد. یعنی یعنی ازمیان ۳۹ مورد تست شده، ۱۱ نفر واقعا سالم هستند و ماشین هم به درستی پیش‌بینی کرده که سالم‌اند. ۲۶ نفر واقعا بیمارند و ماشین هم به درستی پیش‌بینی کرده که بیمارند. از طرفی ۲ نفر واقعا بیمار بوده و ماشین اشتباها آن را سالم پیش‌بینی کرده است. همچنین ۰ نفر واقعا سالم‌اند اما ماشین پیش‌بینی کرده که بیمارند.

خروجی دیگر بدست آمده برای این الگوریتم چنین است:

```
confusion_matrix
[[11  0]
 [ 2 26]]

accuracy: 0.9487179487179487

recall: [1.         0.92857143]

precision: [0.84615385 1.        ]
```

مقدار accuracy درصد درستی و دقت پیش‌بینی‌ها را نشان می‌دهد. برای این روش این مقدار ۰/۹۴ بدست آمده است و این یعنی که ۹۴ درصد پیش بینی‌های انجام شده درست هستند و ۶ درصد نادرستند.

مقدار recall نشان می‌دهد که چند درصد از داده‌های مربوط به هر کلاس را به درستی توانستیم پیش‌بینی کنیم. یعنی در این روش برای کلاس ۰ (سالم)، ۱۰۰درصد و برای کلاس ۱ (بیمار)، ۹۲ درصد از پیش‌بینی‌هایمان درستند.

مقدار precision نشان می‌دهد که چند درصد از از پیش‌بینی‌های انجام شده برای هرکلاس، واقعا در آن کلاس بوده اند. یعنی برای کلاس ۰ (سالم)، ۸۴ درصد از پیش‌بینی‌هایی که انجام شده واقعا در کلاس ۰ (سالم) بوده‌اند. همچنین برای کلاس ۱ (بیمار)، حدود ۱۰۰ درصد از پیش‌بینی‌هایی که انجام شده واقعا در کلاس ۱ (بیمار) بوده اند.

- **K-Nearest Neighbor:**

با رسم confusion matrix heat map برای این الگوریتم Fig8 را خواهیم داشت.

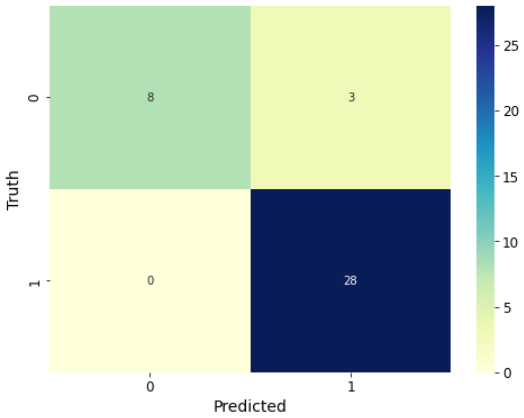


Fig8: K-Nearest Neighbor heat map

این heat map، ماتریس درهم ریختگی یا confusion matrix را نشان می‌دهد که می‌گوید چه تعداد از کل پیش‌بینی‌ها درست بوده است. محور افقی پیش‌بینی‌ها را نشان می‌دهد و محور عمودی واقعیت را. درایه‌های روی قطر تعداد پیش‌بینی‌های درست را نشان می‌دهد. یعنی یعنی ازمیان ۳۹ مورد تست شده، ۸ نفر واقعا سالم هستند و ماشین هم به درستی پیش‌بینی کرده که سالم‌اند. ۲۸ نفر واقعا بیمارند و ماشین هم به درستی پیش‌بینی کرده که بیمارند. از طرفی ۰ نفر واقعا بیمار بوده و ماشین اشتباها آن را سالم پیش‌بینی کرده است. همچنین ۳ نفر واقعا سالم‌اند اما ماشین پیش‌بینی کرده که بیمارند.

خروجی دیگر بدست آمده برای این الگوریتم چنین است:

```
confusion_matrix
[[ 8  3]
 [ 0 28]]

accuracy: 0.9230769230769231

recall: [0.72727273    1.        ]

precision: [1.         0.90322581]
```

مقدار accuracy درصد درستی و دقت پیش‌بینی‌ها را نشان می‌دهد. برای این روش این مقدار ۰/۹۲ بدست آمده است و این یعنی که ۹۲ درصد پیش بینی‌های انجام شده درست هستند و ۸ درصد نادرستند.

مقدار recall نشان می‌دهد که چند درصد از داده‌های مربوط به هر کلاس را به درستی توانستیم پیش‌بینی کنیم. یعنی در این روش برای کلاس ۰ (سالم)، ۷۲ درصد و برای کلاس ۱ (بیمار)، ۱۰۰ درصد از پیش‌بینی‌هایمان درستند.

مقدار precision نشان می‌دهد که چند درصد از از پیش‌بینی‌های انجام شده برای هرکلاس، واقعا در آن کلاس بوده اند. یعنی برای کلاس ۰ (سالم)، ۱۰۰ درصد از پیش‌بینی‌هایی که انجام شده واقعا در کلاس ۰ (سالم) بوده‌اند. همچنین برای کلاس ۱ (بیمار)، حدود ۹۰ درصد از پیش‌بینی‌هایی که انجام شده واقعا در کلاس ۱ (بیمار) بوده اند.

- **Naïve Bayes:**

با رسم confusion matrix heat map برای این الگوریتم Fig9 را خواهیم داشت.

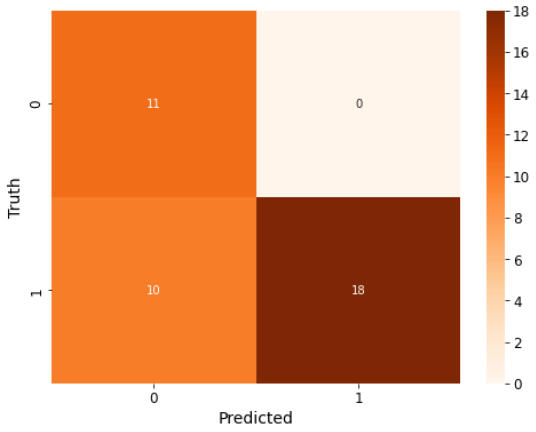


Fig9: Naïve Bayes heat map

این heat map، ماتریس درهم ریختگی یا confusion matrix را نشان می‌دهد که می‌گوید چه تعداد از کل پیش‌بینی‌ها درست بوده است. محور افقی پیش‌بینی‌ها را نشان می‌دهد و محور عمودی واقعیت را. درایه‌های روی قطر تعداد پیش‌بینی‌های درست را نشان می‌دهد. یعنی یعنی ازمیان ۳۹ مورد تست شده، ۱۱ نفر واقعا سالم هستند و ماشین هم به درستی پیش‌بینی کرده که سالم‌اند. ۱۸ نفر واقعا بیمارند و ماشین هم به درستی پیش‌بینی کرده که بیمارند. از طرفی ۱۰ نفر واقعا بیمار بوده و ماشین اشتباها آن را سالم پیش‌بینی کرده است. همچنین ۰ نفر واقعا سالم‌اند اما ماشین پیش‌بینی کرده که بیمارند.

خروجی دیگر بدست آمده برای این الگوریتم چنین است:

```
confusion_matrix
[[11  0]
 [10 18]]

accuracy: 0.7435897435897436

recall: [1.         0.64285714]

precision: [0.52380952 1.        ]
```

مقدار accuracy درصد درستی و دقت پیش‌بینی‌ها را نشان می‌دهد. برای این روش این مقدار ۰/۷۴ بدست آمده است و این یعنی که ۷۴ درصد پیش بینی‌های انجام شده درست هستند و ۲۶ درصد نادرستند.

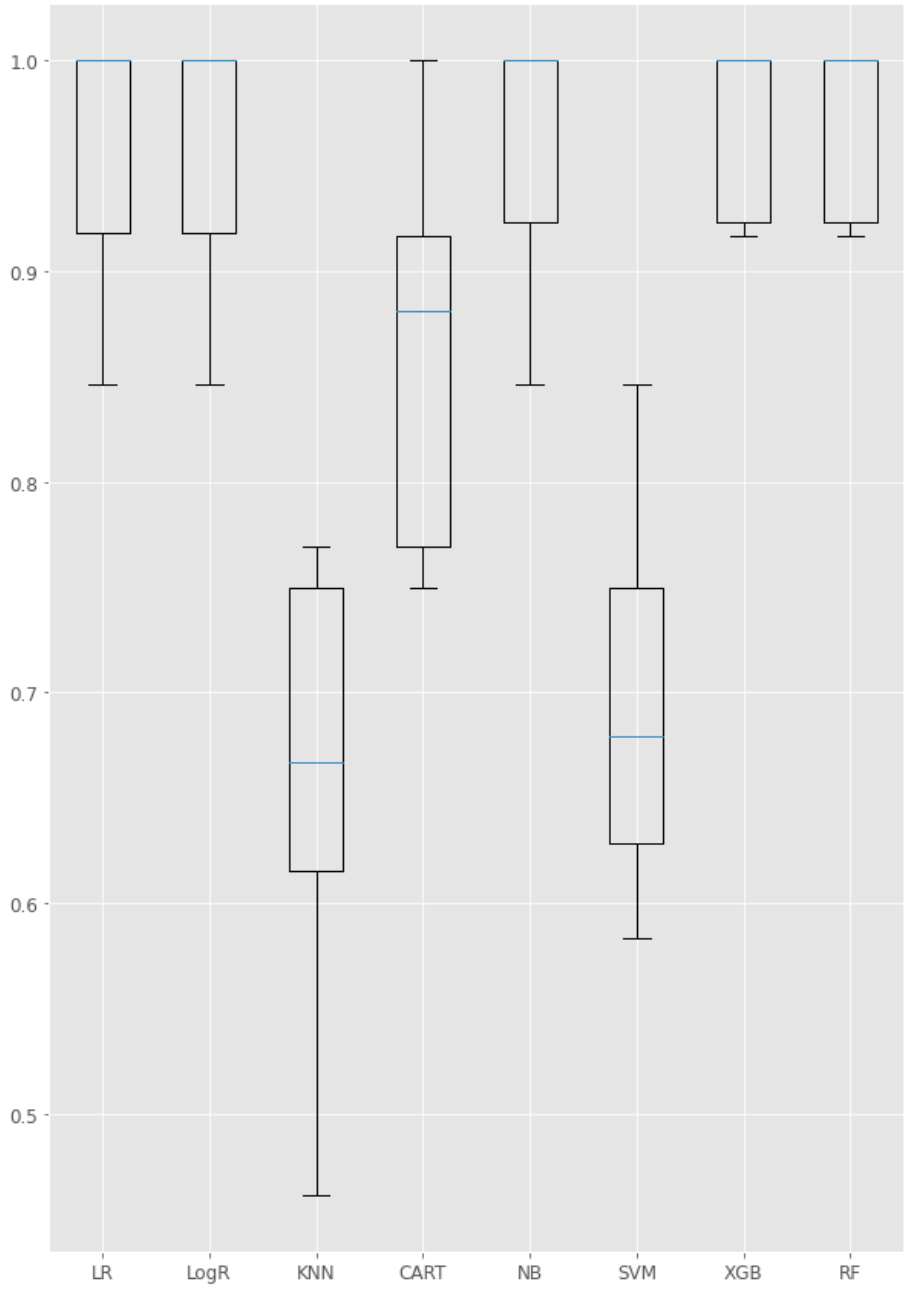
مقدار recall نشان می‌دهد که چند درصد از داده‌های مربوط به هر کلاس را به درستی توانستیم پیش‌بینی کنیم. یعنی در این روش برای کلاس ۰ (سالم)، ۱۰۰درصد و برای کلاس ۱ (بیمار)، ۶۴ درصد از پیش‌بینی‌هایمان درستند.

مقدار precision نشان می‌دهد که چند درصد از از پیش‌بینی‌های انجام شده برای هرکلاس، واقعا در آن کلاس بوده اند. یعنی برای کلاس ۰ (سالم)، ۵۲ درصد از پیش‌بینی‌هایی که انجام شده واقعا در کلاس ۰ (سالم) بوده‌اند. همچنین برای کلاس ۱ (بیمار)، حدود ۱۰۰ درصد از پیش‌بینی‌هایی که انجام شده واقعا در کلاس ۱ (بیمار) بوده اند.

جمع‌بندی نتایج:

هرکدام از این الگوریتم‌ها مزایا و معایبی دارند و برای حل هر مسئله با توجه به ویژگی های آن مسئله می‌بایست بهینه ترین روش با کمترین خطا را انتخاب کنیم. اما آنچه در این مسئله، با مقایسه accuracy واضح است، این است که دقت پیش‌بینی در الگوریتم Extreme Gradient Boosting بیش از دیگر الگوریتم‌ها است.

```
Linear Regression Accuracy :
 0.6634994862742398

Logistic Regression Accuracy :
 0.9743589743589743

Decision Tree Accuracy :
 0.9230769230769231

Support Vector Machine Accuracy :
 0.9487179487179487

Random Forrest Accuracy :
 0.9487179487179487

XGBClassifier Accuracy :
 1.0

Neural Network Accuracy :
2/2 [==============================] - 0s 4ms/step - loss: 0.1659 - accuracy: 0.9487
 [0.1658591479063034, 0.9487179517745972]

KNN Accuracy :
 0.9230769230769231

Naive Bayes Accuracy :
 0.7435897435897436
```


Algorithms Comparision
1.0
0.9
0.8
0.7
0.6
0.5
LR
LogR
KNN
CART
NB
SVM
XGB
RF