
بسم الله الرحمن الرحیم

دانشکده مهندسی برق و کامپیوتر

سامانهٔ پیش‌بینی ابتلا به دیابت
(مبتنی بر الگوریتم‌های یادگیری ماشین)

نام و نام خانوادگی دانشجو: قوام الدین سلیمانی

استاد راهنما: سرکار خانم دکتر فرشته دهقانی

تابستان ۱۴۰۲

تقدیر و تشکر

تقدیم به همهٔ کسانی که با پاکدلی، پرتوی دانایی و آگاهی را بر تاریکی‌های کانایی می‌افشانند.

فهرست مطالب

فهرست تصاویر

خلاصه (چکیده)

بیان موضوع: دیابت یک بیماری مزمن است که فرد مبتلا، قند خون بالاتر از حد مجاز را دارا ست و این مسئله موجب عوارض و مشکلات جدی در سلامت وی (از جمله برخی نارسایی‌ها، سکته‌ها، آسیب و از کار افتادن اندام ها) می‌شود. لازمه ابتلا و بروز این بیماری، عوامل ژنتیکی و محیطی می‌باشد. لذا در صورت وجود احتمال ابتلا به این بیماری در افراد، می‌توان با تغییر سبک زندگی و کنترل‌های پزشکی، تا حدی از ابتلای به این بیماری در افراد مستعد و محتمل، جلوگیری کرد.

روش تحقیق: روش هایی که بتواند به ما کمک کند تا با دقت مناسبی بتوانیم ابتلای افراد مختلف به بیماری را در آینده را پیش‌بنی کنیم، بسیار حائز اهمیت هستند. یادگیری ماشین و داده‌کاوی با استفاده از داده‌های مختلف که در گذشته جمع آوری شده اند می‌توانند کمک شایانی به ما در این امر داشته باشند. پس از الگوریتم‌های مختلف یادگیری ماشین از جمله رگرسیون لجستیک، جنگل درختان تصادفی و... استفاده کردیم تا دقت هر یک را اندازه گیری کنیم.

داده هایی که برای آموزش مدل‌هایمان استفاده کرده ایم، از مجموعه دادگان بیماران آمریکایی[1] است که در سال ۲۰۰۹ تا ۲۰۱۲ جمع آوری شده بودند. این داده‌ها را با استفاده از روش‌های داده‌کاوی گوناگون، پردازش و سپس مصورسازی کردیم و سپس با الگوریتم‌های مذکور و معیارهای ارزیابی مربوط به آن‌ها، سعی در شناسایی بهترین مدل پیش‌بینی کننده کردیم.

یافته‌ها و نتایج: مدل جنگل درختان تصادفی حاوی بهترین نتایج نسبت به سایر مدل‌ها بود...

...

[1] NHANES

فصل ۱

مقدمه

۱-۱ تاریخچه

دیابت چیست ؟

دیابت، یک بیماری مزمن است و زمانی رخ می‌دهد که بدن انسولین کافی تولید نمی‌کند یا نمی‌تواند به طور موثر از انسولین تولید شده استفاده کند.[۵] انسولین هورمونی است که به تنظیم سطح قند خون کمک می‌کند. هنگامی‌که دیابت به درستی مدیریت نشود، می‌تواند منجر به عوارض جدی سلامتی مانند بیماری‌های قلبی، انواع سکته مغزی و قلبی، نارسایی کلیه، کوری و آسیب عصبی شود. [۱۳] آمار ابتلا و مرگ و میر بسیار بالایی از این بیماری در جهان وجود دارد و متاسفانه روز به روز این آمار افزایش می‌یابد.

طبق آمارها، از هر ۱۰ نفر که به دیابت مبتلا هستند، بیش از ۸ نفر آن‌ها از این مسئله آگاهی ندارند و عدهٔ زیادی از افراد هم به پیش‌دیابت مبتلا هستند.[۶] در پیش‌دیابت، سطح قند خون بالاتر از حد طبیعی است، اما به اندازه کافی برای تشخیص دیابت بالا نیست. پیش‌دیابت خطر ابتلا به دیابت، بیماری قلبی و سکته را افزایش می‌دهد.[۵] اگر پیش‌دیابت در افراد وجود داشته باشد، یک برنامه برای تغییر سبک زندگی، می‌تواند به افراد در جلوگیری از این بیماری کمک کند.[۱۳]

این بیماری سه نوع دارد:[۵]

۱. دیابت نوع اول که معروف به دیابت جوانی است چون افراد با سن کمتر از ۳۰ سال معمولا مبتلا می‌شوند. در این نوع به طور ساده می‌توانیم بگوییم میزان انسولین مورد نیاز که توسط پانکراس بایستی ساخته شود و در خون وجود داشته باشد کافی نیست.

۲. دیابت نوع دوم که به بزرگسالی معروف است و در افراد میانسال و مسن رایج تر است در اثر عدم جذب انسولین موجود در خون توسط سلول‌ها می‌باشد.

۳. دیابت نوع سوم دیابت بارداری است که در خانم‌های باردار به طور موقت اتفاق می‌افتد.

۱-۲ علل ابتلا

در ابتلا به این بیماری بنا به نوع آن و همچنین شرایط ژنتیکی و محیطی افراد مختلف، فاکتورهای متنوعی مطرح است: [۵]

۱-۲-۱ فاکتورهای محیطی

مطابق تحقیقات و بررسی‌های انجام شده از سال‌ها پیش تا کنون، عوامل سبک زندگی چون رژیم غذایی نامناسب، عدم فعالیت بدنی و اضافه وزن (مخصوصا میزان توده بدنی) می‌تواند خطر ابتلا به دیابت را افزایش دهد.[۱۳] همچنین وجود بیماری‌های زمینه‌ای مثلا در پانکراس بین افراد می‌تواند در مبتلا شدن به این بیماری موثر باشد که بنا به تعریف دیابت نوع یک، این عامل مربوط به همین نوع می‌شود.[۵]

۱-۲-۲ عوامل ژنتیکی و غیرمحیطی

برخی از افراد استعداد ژنتیکی برای دیابت دارند، به این معنی که بدن آن‌ها بیشتر در معرض ابتلا به این بیماری است. عواملی مثل جنسیت، نژاد و شاخص هایی خونی مختلف که می‌تواند در اثر بیماری‌های خانوادگی و ارثی دیگری در افراد وجود داشته باشد. مثل برخی ویروس‌ها، وجود کلسترول، چربی و فشار خون و ... [۱۳] [۳] [۴]

۱-۳ راهکارهای پیش‌گیری و استعدادسنجی

مطابق توصیهٔ متخصصین اگر بتوانیم افرادی را که استعداد ابتلا به این بیماری را دارند، شناسایی کنیم و این افراد سبک زندگی و روش هایی خاصی را در پیش بگیرند، می‌توانند از ابتلا به این بیماری پیش گیری کنند.[۳]

۱-۳-۱ مبتلایان به پیش‌دیابت

مطابق توصیه پزشکان، در افرادی که به پیش‌دیابت مبتلا باشند یا سابقه این بیماری در خانواده آن‌ها وجود داشته باشد، به طور پیش فرض باید بر یک سبک زندگی سالم، اهتمام ورزند. در این راستا می‌توان به موارد ذیل اشاره کرد:[۴]

- حفظ رژیم غذایی غنی از فیبر مثل انواع میوه‌ها و سبزیجات و کاهش مصرف غذاهای شور، چرب و شیرین
- ورزش منظم
- استفاده از برخی داروها مطابق تجویز پزشک

۱-۳-۲ سایر افراد جامعه

مطابق آمارها، سالانه بخش دیگری از افراد جامعه که از دستهٔ قبلی سوا بوده اند، به بیماری دیابت مبتلا می‌شوند.[۳] در این جا با تحلیل برخی فاکتورهای سلامتی می‌توان پیش‌بنی کرد که آیا این افراد ممکن است با ادامهٔ سبک زندگی کنونی، در آینده به این بیماری دچار شوند و آیا بهتر است با تغییر سبک زندگی خود از ابتلا به این بیماری جلوگیری کنند یا نه ؟

در این زمینه تحقیقات آماری و بررسی‌های مختلفی انجام شده تا بتوانیم با اندازه گیری برخی فاکتورهای کمی‌و کیفی در افراد، مسئلهٔ استعداد در ابتلا به این بیماری را در آن‌ها بررسی کنیم. چالشی که در این زمینه وجود دارد این است که بسیاری از داده‌های غیرخطی و غیراستاندارد پزشکی با ارتباطات و ساختارهای پیچیده وجود دارند که این بررسی‌ها را دشوار می‌سازد.[۶]

۱-۴ انگیزه و اهداف انجام این پژوهش

در راستای همه موارد مطرح شده در بخش‌های قبلی، بر آن شدیم تا تحقیق کنیم با توجه به امکانات و امروزی و دسترسی به مقالات و منابع گوناگون و همچنین توسعهٔ ابزارهای مبتنی بر یادگیری ماشین و هوش مصنوعی، در صدد یافتن بهترین راهکارها برای نجات جان انسان‌های بیشتر با درنظرگیری مناسب‌ترین الگوریتم‌ها باشیم.

اگر بتوانیم افرادی را که احتمال ابتدا به دیابت در آینده برای آنان زیاد است را شناسایی کنیم می‌توانیم با ارائه برنامه‌های پزشکی مناسب از ابتلای افراد به بیماری مذکور جلوگیری کنیم.

۱-۵ موارد انجام شده

در ابتدا مقالات مختلفی را مطالعه کردم و درمورد الگوریتم هایی که مورد بررسی قرار دادم از جمله رگرسیون لجستیک، درخت تصمیم، جنگل درختان تصادفی و بیز ساده، اطلاعات زیادی کسب کردم.

پس از یافتن مجموعه دادگان نمونه که مربوط به اطلاعات بیماران آمریکایی در سال‌های ۲۰۰۹ تا ۲۰۱۲ بوده است، با به کار گیری کتابخانه‌های مختلف پایتون از جمله scikit learn، Pandas، seaborn، numpy و ... عملیات‌های گوناگونی بر روی داده‌ها انجام شد. از جمله: تمیز کردن دادگان، مقیاس بندی و عملیات توزیع مختلف، تصویر سازی و در نهایت مدلسازی و دستیابی به نتایج نهایی که در نهایت الگوریتم جنگل درختان تصادفی با دقت **فلان** بهترین عملکرد را بین باقی الگوریتم‌ها در این پیش‌بینی، شناخته شد.

فصل ۲

ادبیات پژوهش

۲-۱ مقدمه

روش‌های نظارت شده‌ای مانند طبقه بندی و تخمین تلاش می‌کنند تا رابطه‌ای میان صفات خاصهٔ ورودی (که گاه متغیرهای مستقل نامیده می‌شوند) را با یک یا چند صفت خاصهٔ هدف (که گاه متغیر وابسته نامیده می‌شود) کشف کنند. در نهایت این رابطه با یک ساختار به عنوان مدل نمایش داده می‌شود. [۱]

. . . . .

۲-۲ روش‌های داده‌کاوی

۲-۲-۱ طبقه‌بندی و خوشه‌بندی

توضیح درمورد یادگیری ماشین و خوشه بندی و طبقه بندی و ...

۲-۲-۲ بیش برازش و کم برازش در فرایند ها

۲-۳ معرفی الگوریتم‌های طبقه بندی

۲-۳-۱ رگرسیون لجستیک [۱]

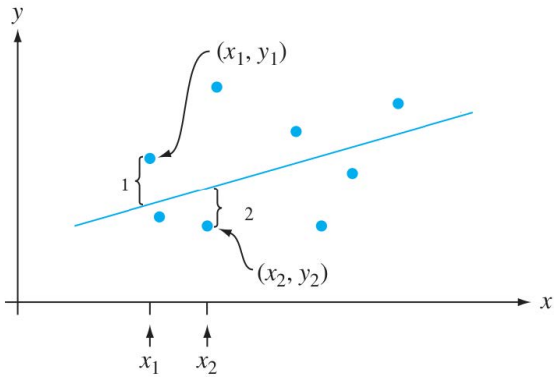
خوب است در ابتدا مروری بر رگرسیون خطی [۲] داشته باشیم تا در ادامه بتوانیم رگرسیون لجستیک را بهتر درک کنیم.

رگرسیون خطی:

در بسیاری از بررسی‌های آماری، لازم است یک متغیر وابسته را از روی یک یا چند متغیر مستقل

[۱] Logistic Regression
[۲] Linear Regression

پیش‌بینی کنیم که اصطلاحا به آن رگرسیون یا برگشت می‌گوییم.[۲] برای مثلا میزان ساعت مطالعه یک متغیر مستقل است و نمره اخذ شده در درسی متغیری وابسته است و بین این دو رابطه وجود دارد. سپس نمونه‌ای از جمعیت را در نظر گرفته و در آن مقدارهای X_1 تا X_n در متغیر مستقل خود مقابل مقادیر نظیر در متغیر وابسته از Y_1 تا Y_n قرار می‌دهیم.[۲] سپس آن‌ها را مثل یک نمودار در صفحه مختصات به یکدیگر متصل کرده که به آن نمودار پراکندگی گوییم.[۲]



شکل ۲-۱: نمودار پراکندگی

حالا می‌توان خطی را در این صفحه مختصات درنظر گرفت که تا حد زیادی منطبق بر نقاط باشد که در واقع یک نمودار پیش‌بینی کننده Y بر مبنای X است که به آن معدله رگرسیون Y بر روی X گویند.[۲] حالا رابطهٔ این نقاط و منحنی را با $\mu_{Y|x} = E(Y|x) = \alpha + \beta x$ مشخص می‌کنیم و α و β پارامتر هایی هستند که باید مقدار دهی شوند تا خط بر نقاط منطبق باشد.[۲] مسئله‌ای که در اینجا مطرح است این است که ممکن است خط ما بر نقاط مختلف منطبق نشود. لذا اینجا باید حالت بهینه‌ای را در نظر گرفت حداقل مقدار خطا (یا بهتر بگوییم اختلاف) در مجموع داشته باشیم که روش حداقل مربعات برای یافتن میزان بهینهٔ α و β که دارای بیشترین انطباق و کمترین خطا باشند برای ما کمک کننده است.[۲] بر اساس همین روش، با معادلات زیر به مقادیر بهینهٔ α و β دست می‌یابیم.[۲]

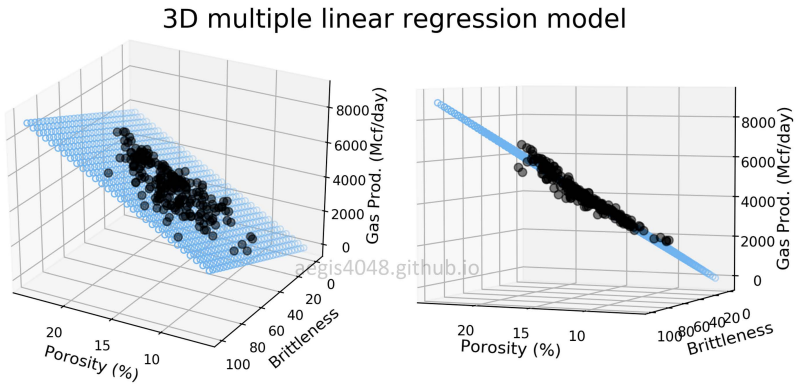
$$\alpha = \bar{y} - \beta\bar{x} \qquad (۱-۲)$$

$$\beta = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \qquad (۲-۲)$$

که در روابط فوق $\bar{x}$ و $\bar{y}$ میانگین y و x هستند. با تعمیم این روابط و اصول بیان شده می‌توان حالتی را درنظر گرفت که چندین متغیر مستقل داریم.(مثلا ۲تا) که این مدل به رگرسیون خطی چندگانه معروف است و در آنجا نمودار ما حالت فضایی پیدا خواهد کرد و با رابطهٔ زیر می‌توانیم آن را بیان کنیم.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ + \beta_p x_{ip} + \epsilon \qquad i = 1, ..., n \qquad (۳-۲)$$

که در این رابطه، p ابعاد ما می‌باشد. [۱۹]



شکل ۲-۲: نمودار رگرسیون خطی چندگانه در فضا

در کل می‌توانیم بگوییم روش‌های رگرسیون زمانی مناسب است که مقادیر مستقل در مجموعهٔ داده‌ها به کلاس هایمان (به بیان دیگر طبقه بندی‌ها) وابستگی داشته باشند. ضریب همبستگی خطی که با رابطهٔ ۲-۴ می‌شود میزان این وابستگی را برای ما نشان می‌دهد.

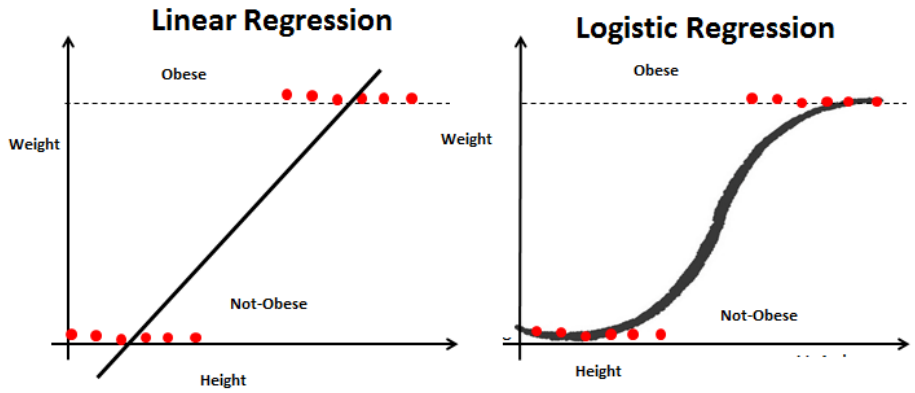
$$p(X,Y) = corr(X,Y) = \frac{Cov(x,y)}{(Var(x)Var(y))^{\frac{1}{2}}} \qquad (۲-۴)$$

صورت کسر کواریانس x و y است و مخرج واریانس x و واریانس y می‌باشد. حاصل مقداری است از ۱- تا ۱ که میزان وابستگی مستقیم یا معکوس را نشان می‌دهد [۱] و در صورت نبود میزان وابستگی مقدار ۰ است. [۲]

رگرسیون لجستیک:

حالا که با مفاهیم ابتدایی رگرسیون آشنایی پیدا کردیم به بیان نوع دیگری از آن به نام رگرسیون لجستیک می‌پردازیم.

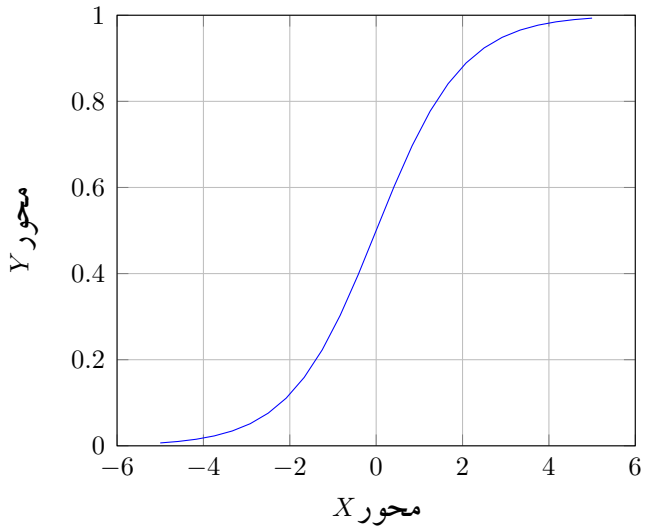
در این مدل به جای اینکه مقدار عددی برای متغیر وابسته تعریف شود، بر اساس احتمال متغیرهای دودویی را برای پیشگویی در نظر داریم [۱] که در ادامه این مسئله را بیشتر توضیح می‌دهیم. مثلا کار ما، فاکتور BMI در افراد یک متغیر مستقل است و بر ابتلا به دیابت در افراد موثر بوده؛ همچنین در اینجا متغییر وابستهٔ ما، همان دیابت گرفتن یا نگرفتن فرد می‌باشد که با ۰ و ۱ آن را درنظر می‌گیریم. پس این مدل برای مواردی استفاده می‌شود که حالت کلاس بندی برای متغیرهایمان داریم.[۸] [۹] ضمنا در حالت کلاس بندی، نمی‌توانیم حالت ۰ یا ۱ را به عنوان اعدادی در حالت رگرسیون خطی منظور کنیم؛ چرا که ۰ و ۱ را به عنوان مقادیر عدد درنظر گرفته می‌شود و امکان نمایش کلاس‌های گوناگون در در یک نمودار خطی مشخص وجود ندارد.[۹] اگر به شکل ۲-۳ دقت کنید، متوجه می‌شوید در حالت رگرسیون خطی برخی از نمونه‌ها در کلاس مربوطه قرار نگرفته اند.



شکل ۲-۳: مقایسه رگرسیون لجستیک و خطی

علت این امر مشخص است. زیرا زمانی که کلاس‌های گوناگون داریم، ساختار استدلالی که در الگوریتم رگرسیون خطی مطرح است نمی‌تواند طبقه بندی صحیحی برای ما انجام دهد. نهایتا چاره این است که از یک نمودار منحنی شکل برای فشرده کردن نتایج بین ۰ تا ۱ استفاده کنیم [۹]: رابطه ۲-۵ و شکل ۲-۴

$$logistic(x) = \frac{1}{1 + exp^{-x}} \qquad (۵-۲)$$



شکل ۲-۴: نمودار تابع لجستیک

سپس اگر سمت راست رابطهٔ ۲-۳ را در جای X در معادلهٔ ۲-۵ (که آن را تابع سیگماوار[۳] می‌نامند) قرار دهیم، فقط مقادیر ۰ و ۱ را به ما می‌دهد.

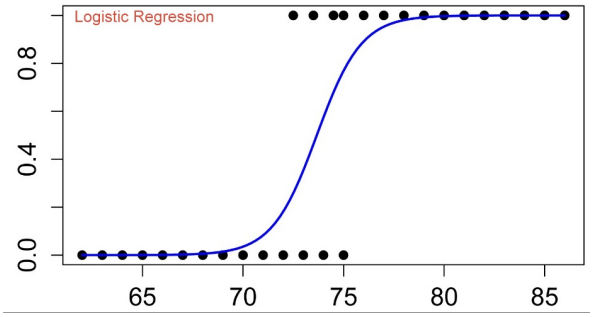
حال در این مدل، چون قرار است هر نمونه به یک کلاس تعلق یابد، بر اساس روابط مطرح شده در

[۳] Sigmoid/ Logistic Function

این بخش و تعمیم خواص آماری (واریانس)، می‌دانیم وزن ویژگی هایمان عاملی اثر گذار تعیین مرز جداسازی در نمودار ما خواهد بود. لذا نهایتا به رابطهٔ زیر می‌رسیم که بر اساس احتمال قرار گیری هر نمونه در هر کلاس برای ما کلاس‌بندی را انجام می‌دهد [۸][۹]:

$$g(x) = ln(\frac{p(x)}{1-P(x)}) = \frac{\frac{e^{b_0+b_1x}}{1+e^{b_0+b_1x}}}{1-\frac{e^{b0+b_1x}}{1+e^{b_0+b_1x}}} = ln(e^{b_0+b_1x}) = b_0 + b_1x \qquad (۶-۲)$$

در رابطهٔ فوق، کاری که انجام می‌شود به این صورت است که احتمال قرار گیری یک نمونه در یک کلاس نسبت به حالت دیگر سنجیده می‌شود و درصورتی که مثلا با احتمال بالای ۵۰ درصد در طبقه ۱ قرار می‌گیرد، این کار صورت می‌پذیرد و در غیر این صورت به کلاس ۰ متعلق است. (می دانیم که احتمال دقیقا ۵۰ درصد روی نقطعه (۰،۰) قرار دارد.) [۸][۹] به بیان دیگر می‌توانیم بگوییم، Logit پاسخِ Y، ترکیب خطی پیش‌بینی‌کننده‌ها (یا همان X) است. [۶]



شکل ۲-۵: مدل Logist Regression

۲-۳-۲ درخت تصمیم

۲-۳-۳ جنگل درختان تصادفی

این الگوریتم از ساختار درختان تصمیم استفاده می‌کند. ساختار درختان تصمیم...

حال می‌توانیم بگوییم این الگوریتم، با ایجاد چندین درخت تصمیم مختلف از داده هایمان برای ما تصمیم گیری را انجام می‌دهد. ممکن است بپرسید که خب، آیا مثلا ۴ درخت تصمیم مختلف نتایج یکسانی دارند؟ با توجه به ساختار درختان تصمیم جواب قطعا خیر است. پس در اینجا از جواب‌ها رای گیری میشود. یعنی در نمونه‌ای ۳ درخت حکم می‌کنند که فرد به دیابت مبتلا خواهد شد و ۱ درخت حکم می‌کند که این فرد به دیابت مبتلا نخواهد شد. لذا اکثریت آراء بر مبتلا شدن فرد اتفاق نظر دارند. پس نتیجهٔ آن پیش‌بینی، ابتلا شدن فرد است.

ادامه دارد...

۲-۳-۴ AdaBoost

این الگوریتم یک فرض ساده در نظر دارد و آن یادگیری گروهی است.[۱۲] فرض کنید کسانی که در مجموعه داده‌های آموزشی دچار اشتباه در تشخیص ابتلا به بیماری دیابت شدند، از مجموعه

بقیه داده‌های آموزشی جدا می‌شوند و در یک طبقه بندی جدید مجددا مورد ارزیابی قرار می‌گیرند؛ همانطور که در واقعیت ممکن است هرپزشکی معیارهای مختلفی را برای تشخیص بیماری مراجعه کننده اش داشته باشد و هر پزشکی نمی‌تواند همهٔ افراد را درست تشخیص دهد پس اگر یک تیم پزشکی داشته باشیم نظرات جمعی می‌توانند بیماران بیشتری را به درستی شناسایی کنند.[۱۰]

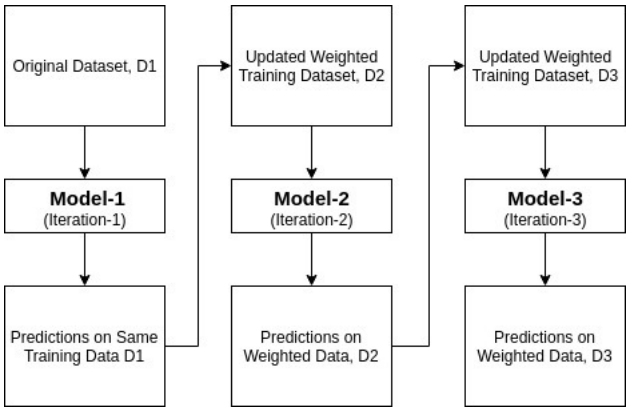
این روند جداسازی داده‌های آموزشی و امتیاز دهی در تشخیص توسط طبقه بندی کننده‌های مختلف درون درختان تصمیم ضعیف، مکررا تکرار می‌شود تا درنهایت طبقه بندی‌های متعددی داشته باشیم که هر کدام بر اساس میزان سرآمدی در تست‌های مختلف، امتیازات مختلفی دریافت کنند.[۱۲] بعد از ساخته شدن مدل، حالا می‌توانیم به نسبت امتیاز هر طبقه بندی کننده داده‌های دریافتی را به صورت شانسی بین هرکدام تقسیم کنیم تا پیش‌بینی صورت گیرد و این مسئله موجب تقویت سنجش می‌گردد. به همین دلیل به آن الگوریتم Adaptive Boosting یعنی تقویت کنندهٔ تطبیقی گویند. [۱۰] [۱۲]

اگر از دید ریاضی الگوریتم را بررسی کنیم، برای محاسبهٔ میزان خطای مدل، M_i وزن هر یک از تاپل‌های D_i را که M_i اشتباه طبقه بندی کرد، جمع می‌کنیم.

$$error(M_i) = \sum_{j=1}^{d} w_i \times err(X_j) \qquad (۷-۲)$$

حالا اگر طبقه بندی کننده X_j اشتباه کند، میزان ارور برای آن ۱ اندازه گیری می‌شود و در غیر این صورت ۰ است. اگر یک طبقه بندی کننده آن قدر ضعیف باشد که خطایش از 0.5 بیشتر شود آن را دیگر در نظر نمی‌گیریم.[۱۰] سپس یک مجموعهٔ دادهٔ جدید به نام D_i تولید می‌کنیم و از آن یک M_i جدید استخراج می‌کنیم و دوباره این روند را ادامه می‌دهیم. هر چه میزان خطای طبقه بندی کننده کمتر باشد، دقیق تر است و بنابراین، وزن آن برای انتخاب شدن باید بیشتر باشد. از رابطه ۲-۸ برای محاسبهٔ وزن هر طبقه بندی استفاده می‌شود.[۱۰]

$$log(\frac{1 - error(M_i)}{error(M_i)}) \qquad (۸-۲)$$



شکل ۲-۶: الگوریتم AdaBoost: در اینجا مراحل ذکر شده به صورت متناوب تکرار می‌شود.[۱۶]

Naive Bayes ۵-۳-۲

۴-۲ داده‌ها و مصورسازی

داده هایی که از آن‌ها استفاده کردیم، برگرفته از یک برنامه مطالعاتی به نام NHANES بوده که جهت بررسی سلامتی کودکان و بزرگسالان از سوی CDC[۴] تدوین شده است.

برنامه NHANES در اوایل دهه ۱۹۶۰ آغاز شد و به صورت مجموعه‌ای از نظرسنجی‌ها با تمرکز بر گروه‌های مختلف جمعیتی یا موضوعات بهداشتی انجام شده است. در سال ۱۹۹۹، این نظرسنجی به یک برنامه مستمر تبدیل شد که تمرکز در حال تغییری بر روی انواع اندازه گیری‌های سلامت و تغذیه برای رفع نیازهای نوظهور دارد.[۱۴]

مصاحبه NHANES شامل سوالات جمعیت شناختی، اجتماعی –اقتصادی، رژیم غذایی و سلامتی است. جزٔ معاینه شامل اندازه گیری‌های پزشکی، دندانی و فیزیولوژیکی و همچنین تست‌های آزمایشگاهی است که توسط پرسنل پزشکی بسیار آموزش دیده انجام می‌شود.[۱۴]

۱-۴-۲ داده‌های گم‌شده

داده‌های گم‌شده چیست؟

راه حل‌های داده‌های گم شده

- حذف ردیف‌های حامل داده‌های گم شده
- جایگزینی با متوسط‌های آماری

۲-۴-۲ افراز داده‌ها

بخش کردن داده‌ها یا افراز [۵]

۵-۲ استاندارد سازی داده‌ها

برای ادامهٔ مراحل، لازم است کارهای بیشتری را روی داده‌های خود انجام دهیم. از جمله ایجاد متغییرهای ساختگی (دو دویی) و مقایس بندی

۱-۵-۲ متغیرهای ساختگی (دودویی)

۲-۵-۲ مقیاس‌بندی

برخی از الگوریتم‌ها نیاز دارند که داده‌ها قبل از پیاده سازی مؤثر الگوریتم نرمال‌سازی شوند. برای نرمال سازی یک متغیر، میانگین را از هر مقدار کم می‌کنیم و سپس بر انحراف استاندارد تقسیم می‌کنیم.

[۴] US Centers for Disease Control and Prevention
[۵] Data Splitting

این عملیات گاهی اوقات استانداردسازی نیز نامیده می‌شود.[۱۱]

متغیرهایی که در مقیاس‌های مختلف اندازه گیری می‌شوند به طور یکسان در برازش مدل و تابع آموخته شده مدل نقش ندارند و ممکن است منجر به ایجاد یک سوگیری شوند.

https://towardsdatascience.com/how-and-why-to-standardize-your-data-996926c2c832
https://www.geeksforgeeks.org/normalization-vs-standardization/

۲-۵-۳ Cross validation

۲-۵-۴ مصورسازی

تعریف: به طور ساده می‌توانیم بگوییم زمانی که داده هایمان را به صورت انواع نمودارها، نقشه‌ها و شکل‌های مختلف بصری دربیاوریم تا نتیجه‌گیری و تحلیل آن‌ها توسط مغز جهت شناسایی الگوها و نقاط پرت در داده آسان تر شود، این کار انجام می‌گیرد. [۱۵]

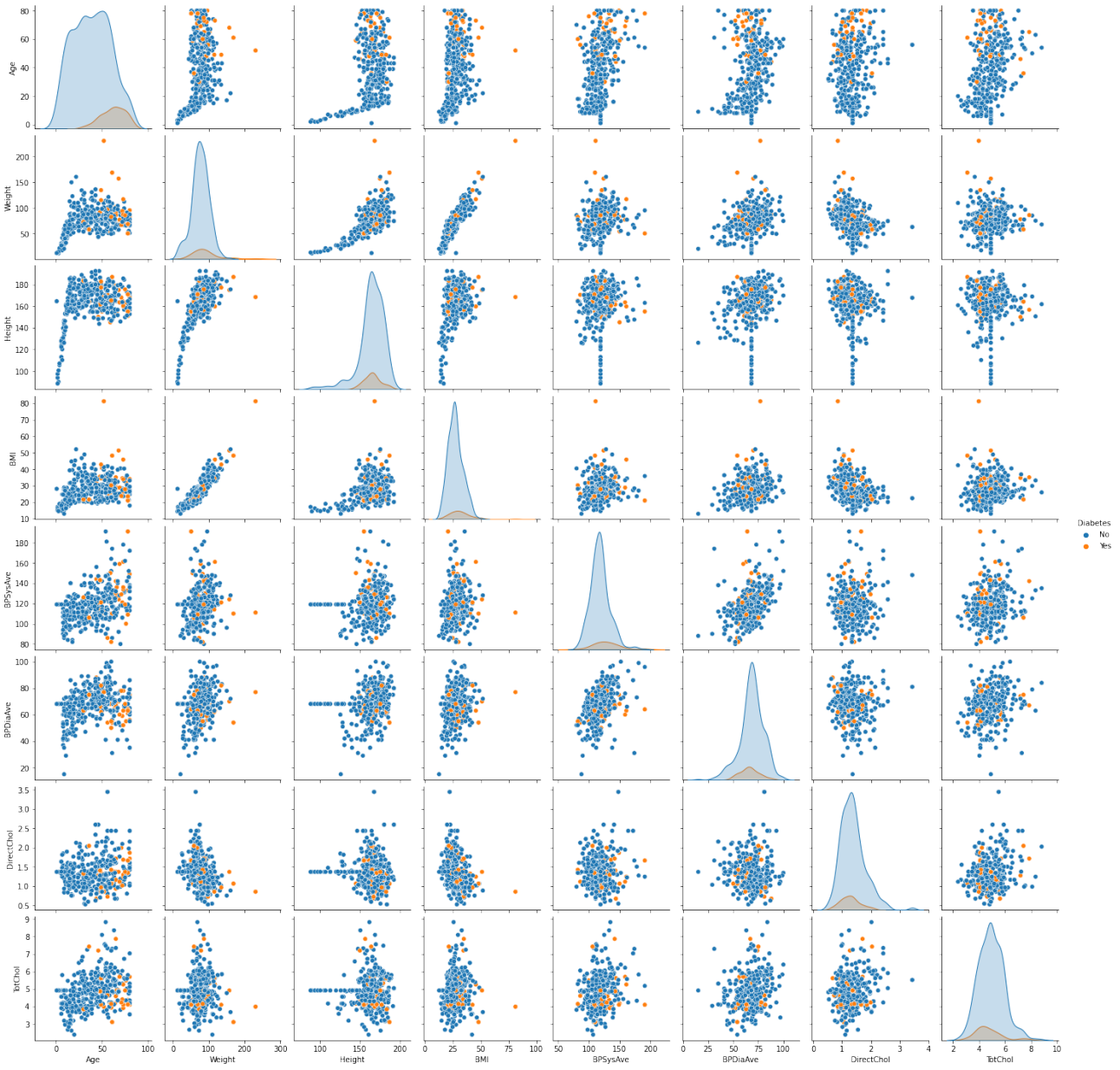
اهمیت: یک تصویر هزاران برابر بیشتر از کلمات ارزش دارد.

جملهٔ فوق، به نوعی اهمیت مصورسازی را برای ما نمایان می‌کند. درواقع ما با انجام این‌کار، درک بهتر و سریع‌تری نسبت به داده‌های عظیم خود خواهیم داشت. قابل ذکر است که Dataset مورد استفاده ما در اینجا، حدود ۱۰۰۰۰ رکورد را در خود جای داده است.

ضمنا در کنار این موارد، با مصور سازی داده‌ها دیگر نیازی به توضیحات اضافه نخواهیم داشت و بسیاری از افراد عادی قادر به درک موضوع مطرح شده خواهد بود. [۱۵]

۲-۵-۵ انواع نمودارها و داده‌های اجمالی

Pairplot



شکل ۲-۷: Pairplot

در شکل ۲-۷ نمودار Pairplot را مشاهده می‌کنیم.

در این نمودار، نقاط نارنجی رنگ نشان دهندهٔ افراد مبتلا به دیابت است و نقاط آبی رنگ نشان دهندهٔ افرادی بدون ابتلا به این بیماری است.

ادامه دارد....

۲-۶ نتیجه‌گیری

فصل ۳

کارهای پیشین

۳-۱ مقدمه

در انجام پژوهش‌های گوناگون بررسی و مقایسه روش‌های مختلف در روند توسعه و تحقیق مجدد، نکته بسیار مهمی است و می‌تواند اشکالات کارهای پیشین را برطرف نمود و در زمینه موارد به روز آن‌ها را به کار گرفت. همچنین ایده‌ها و نکاتی در هر مقاله ذکر شده است که می‌تواند رهنمودهای مفید برای کارهای آتی طلقی شوند.

۳-۲ مقاله ۱

مقاله [۷] در سال ۲۰۱۸ با استفاده از داده‌های بیماران هندی[۱] روند بررسی را بر روی سه الگوریتم درخت تصمیم، SVM و بیز ساده انجام داده و نتایج بیانگر این بوده که الگوریتم بیز ساده ٪۷۶ به عنوان موثر ترین الگوریتم در این بررسی درنظر گرفته شده است.

نکته قابل ملاحظه در این مقاله این است که از الگوریتم SVM هم برای پیش‌بینی استفاده شده است که معمولا برای دادگان با مقادیر زیاد روش مناسبی است. [۱۷]

۳-۳ مقاله ۲

در مقالهٔ اشاره شده [۶] در سال ۲۰۲۰ روش‌های مختلفی برای پیش‌بینی ابتلا به دیابت انجام شده و ترکیب جنگل درختان تصادفی و رگرسیون لجستیک در اعتباری سنجی متقابل K-Fold "مکرر" با مقدار K=۱۰ بهترین نتیجه طبقه بندی را با دقت ٪۹۴ حاصل کرده است. در این مقاله روش‌های بیز ساده و AdaBoost هم مورد استفاده قرار گرفتند که نتایج آن‌ها داری دقت مناسبی نبوده است.

ضمنا در این مقاله، از همان مجموعه دادگانی استفاده شده که در همین پایان نامه مورد استفاده قرار گرفته است.

[۱] PIDD

۳-۴ نتیجه‌گیری

در کل با مطالعه مقالات اساسی که در گذشته تالیف شده بودند، سعی شد روش‌های مناسب به کار گرفته شوند و در روند توسعه استفاده شوند که از جمله می‌توان به الگوریتم‌های بیز ساده، جنگل درختان تصادفی و الگوریتم درخت تصمیم و رگرسیون لجستیک اشاره کرد.

فصل ۴

روش‌ها و نتایج

۴-۱ مقدمه

در این بخش بررسی می‌کنیم که الگوریتم‌های گوناگونی که برای ساخت مدل‌های گوناگون استفاده کردیم تا چه حد می‌تواند قابل اعتماد واقع شود و میزان خطا در هرکدام چقدر است. چرا که ما در این پروژه، کار پیش‌بینی انجام دادیم و داده‌های آزمون به ما نشان داده که برخی از ردیف‌های به صورت اشتباه پیش‌بینی شده اند. برای انتخاب بهترین مدل روشی که انجام می‌شود این است که میزان خطاها را با روش‌های گوناگون به دست آوریم و سپس الگوریتمی که بیشترین صحت را ارائه داده به عنوان بهترین مدل برگزینیم. [۱۰]

همچنین برای بخش دیگری از این سنجش می‌توانیم نمودارهای گوناگونی را رسم کنیم از جمله ROC که می‌تواند این میزان دقت را به صورت بصری به ما نمایش دهد.

۴-۲ روش پیشنهادی

۴-۲-۱ اندازه گیری میزان خطای دقت

به دلیل پدید آمدن مشکلاتی از قبیل بیش برازش و کم برازش داده‌ها که در بخش‌های قبلی به آن اشاره شد مدل‌های ما دچار اشتباهاتی در پیش‌بینی‌ها میشود. در این بخش به بیان انواع خطاها و روابطی که برای سنجش آن‌ها به کار گرفته ایم می‌پردازیم. [۱۰]

ابتدا به معرفی ۴ تاپل مختلف می‌پردازیم که بعدا از آن‌ها در محاسبهٔ میزان خطا استفاده خواهیم کرد [۱۰]:

- **TP**[۱]: این تاپل موارد مثبت واقعی را علامت گذاری می‌کند. مثلا در این پروژه مواردی که احتمال ابتلا به دیابت در آن‌ها مثبت پیش‌بینی شده و در دادگان هم مثبت بوده در این دسته قرار می گیرد.
- **TN**[۲]: این تاپل موارد منفی واقعی را علامت گذاری می‌کند. مثلا در این پروژه مواردی که احتمال

[۱] True positives
[۲] True negatives

ابتلا به دیابت در آن‌ها منفی پیش‌بینی شده و در دادگان هم منفی بوده در این دسته قرار می‌گیرد.

- **FP**[۳]: این تاپل موارد مثبت کاذب را علامت گذاری می‌کند. مثلا در این پروژه مواردی که احتمال ابتلا به دیابت در آن‌ها مثبت پیش‌بینی شده اما در دادگان هم منفی بوده در این دسته قرار می‌گیرد.
- **FN**[۴]: این تاپل موارد منفی کاذب را علامت گذاری می‌کند. مثلا در این پروژه مواردی که احتمال ابتلا به دیابت در آن‌ها منفی پیش‌بینی شده اما در دادگان هم مثبت بوده در این دسته قرار می‌گیرد.
- FPR یا (1 - Specificity):

رابطه ۴-۱ نشان دهنده نسبت افراد سالم شناسایی شده واقعی به کل افراد سالم (افراد اشتباها بیمار تشخیص داده شده به علاوه افراد سالم شناسایی شده واقعی) است.

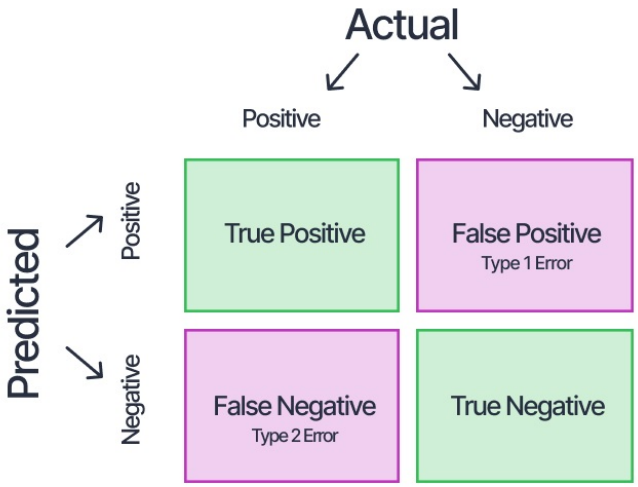
$$1 - Specificity = FPR = \frac{FP}{FP + TN} \qquad (۱-۴)$$

- TPR یا Sensitivity:

رابطه ۴-۲ نشان دهنده نسبت بیماران شناسایی شده واقعی به کل بیماران (افراد اشتباها سالم تشخیص داده شده به علاوه افراد بیمار شناسایی شده واقعی) است.

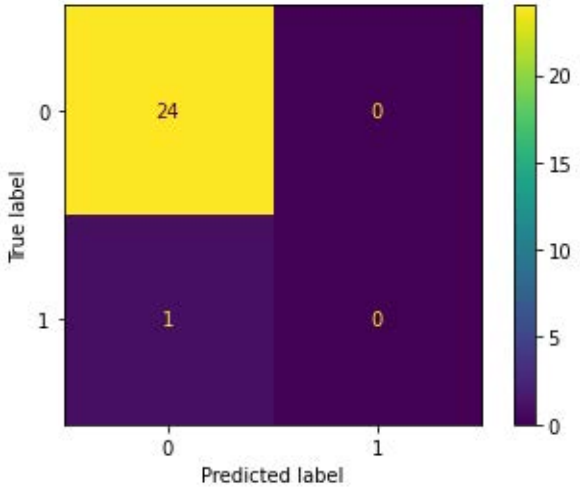
$$Sensitivity(Recall) = TPR = \frac{TP}{TP + FN} \qquad (۲-۴)$$

تعدادی از این مقادیر را در ماتریسی به نام ماتریس آشفتگی[۵] نشان می‌دهند [۱۸] و ساختار آن در شکل ۴-۱ است. همچنین این ماتریس را برای مدل جنگل درختان تصادفی رسم کردیم (شکل ۴-۲) :



شکل ۴-۱: ساختار ماتریس آشفتگی: پارامترهای مذکور در این بخش در این ماتریس قرار گرفته اند.

[۳] False positives
[۴] False negatives
[۵] Confusion



شکل ۴-۲: ساختار ماتریس آشفتگی برای الگوریتم جنگل درختان تصادفی

به طور کلی این ماتریس به ما خلاصه‌ای از درستی نتایج پیش‌بینی هایمان را نشان می‌دهد. [۱۸] حال میزان میزان دقت را با رابطهٔ ۴-۳ محاسبه می‌کنیم.

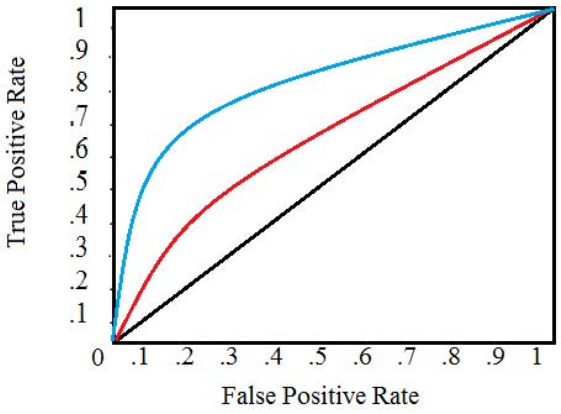
$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad (۴-۳)$$

۴-۲-۲ AUC و ROC

برای اینکه مقادیر پیش‌بینی‌های صحیح و غلط را به صورت یک نمودار نمایش دهیم، از دو شاخصهٔ FPR و TPR که در بخش قبلی آن‌ها را معرفی کردیم استفاده می‌کنیم.[۱۰] ما در اینجا مقادیر مختلف برای هر مرز (اشاره به تابع Sigmid که در بخش‌های قبلی مطرح شد) پیش‌بینی را از ۰ تا ۱ اندازه گیری و در نمودار علامت می‌زنیم.[۱۰] اگر مقدار TPR برای یک مرز برابر با ۱ باشد یعنی در آن مرز مدل خیلی خوب عمل کرده و اگر برابر ۰/۵ باشد یعنی تصادفی عمل شده است. [۱۰] مثلا در یک مرز در نظر می‌گیریم اگر مقدار تابع Sigmid از ۰/۷ بیش تر بود، آن مورد را یک مورد مثبت شناسایی کن. در نمودار ROC [۶] مولفهٔ ی طول‌ها برابر با مقدار FPR است و مولفهٔ عرض‌ها برابر TPR در آستانه‌های گوناگون است.[۱۰]

سطح زیر این نمودار را با AUC[۷] نمایش می‌دهیم که بالا بودن این مساحت بیانگر این است که پیش‌بینی‌های بیشتری در این مدل درست بوده است. [۱۰]

[۶] Receiving Operating Characteristic
[۷] Area Under the Curve



شکل ۴-۳: :۳ شکل کلی نمودار ROC

۴-۲-۳ ابزار Node Red

۴-۲-۴ flask

۴-۳ نتیجه‌گیری

فصل ۵

جمع‌بندی و کارهای آتی

۵-۱ جمع‌بندی

۵-۲ کارهای آتی

در مورد روش های توسعه این سیستم می توانیم به مواردی چون سیستم های یادگیری ماشین آنلاین اشاره کنیم که دادگان همواره با اطلاعات جدید به روزرسانی می شوند و در بازه های مختلف توسط ناظر سیستم بهترین الگوریتم ها بر روی آن ها برای کسب بهترین نتایج اعمال می گردد. ضمن اینکه میتوانیم از اینترنت اشیا و امکاناتی ازین قبیل استفاده کنیم. به تازگی گجت ها و کیت های مخصوصی برای گوشی های هوشمند طراحی شده اند که برای سنجش پارامترهای گوناگون سلامتی مورد استفاده قرار می گیرند و با استفاده از گسترش شبکه های پرسرعت اینترنت نظیر 5G به سرعت می توانیم حجم عظیمی از داده های جدید را دریافت کنیم.

به عنوان یک نمونه ساده و سمبلیک می توان از برد الکترونیکی آردوینو که قابلیت نصب گجت های مختلف و سنسور های گوناگون را فراهم می آورد استفاده کرد و یک سیستم یادگیری ماشین آنلاین را طراحی کرد که با سرور ما (مثلا برای آردوینو NodeRed می باشد) تبادل دارد.

همچنین میتوانیم سامانه های موازی با این سیستم را نیز راه اندازی کرد مثل یک سایت پیش بینی کننده احتمال ابتلا به بیماری دیابت در افراد که هر شخصی با وارد کردن پارامتر های خودش می تواند نسبت به وضعیت سلامتی خود در آینده برآوردی داشته باشد.

۵-۲-۱ راه‌اندازی سامانه ثبت گزارشات دیابت و پیش‌بینی

۵-۲-۲ تولید کیت‌های ثبت نتایج دیابت جدید

پیوست

۱- کد استانداردسازی متون فارسی آمیخته به عبارات انگلیسی

در این کد پایتونی از مبحث RegEx که در درس کامپایلر با آن آشنا شدیم، استفاده کردم. فایل ورودی متن عادی است که پس از انجام عملیات، در فایل خروجی شاهد قرار گرفتن عبارات انگلیسی درون تگ lr خواهیم بود.

```
[.]+?\S

import re
filename='import.txt'
lines=[]
with open(filename) as file:
   lines = [str(line.rstrip()) for line in file]

def myfun(a):
   e=a.group(0)
   e=' \lr{'+e+'}'
   return e

w=[]
for i in lines:
   x =re.findall(r"[a-zA-Z0-9\s]+\b(?=\s[^a-zA-Z0-9]*)", i)
   tempi=re.sub(r"[a-zA-Z0-9\s]+\b(?=\s[^a-zA-Z0-9]*)", myfun, i)
   w.append(tempi)
   print(tempi)

with open(r'export_text.txt', 'w+') as fp:
   for item in w:
     fp.write("%s\n" % item)
print('Done')
```

واژه‌نامه

ب

Data داده‌ها

Data داده‌ها

Data داده‌ها

Data داده‌ها

د

Data جاده‌ها

Data جاده‌ها

Data جاده‌ها

Data جاده‌ها

مراجع

[۱] اسماعیلی.مهدی، مفاهیم و تکنیک‌های داده‌کاوی

[۲] نعمت الهی.نادر، آمار و احتمالات مهندسی

[3] What is diabetes?, Aoife M Egan, Sean F Dinneen

[4] Epidemiology of diabetes,Nita Gandhi Forouh,Nicholas J Wareha

[5] Diabetes Cookbook FOR DUMMIES 3RD EDITION, by Alan L. Rubin, MD with Cait James, MS

[6] Classification and prediction of diabetes disease using machine learning paradigm,Md.Maniruzzaman, Md. Jahanur Rahman, Benojir Ahammed and Md. Menhazul Abedin

[7] Prediction of Diabetes using Classification Algorithms, Deepti Sisodia ,Dilip Singh Sisodia

[8] Logistic Regression,Lynne Connelly

[9] Interpretable Machine Learning,Christoph Molnar

[10] Data Mining Concepts and Techniques ,Jiawei Han,Micheline Kamber,Jian Pei

[11] DATA MINING FOR BUSINESS ANALYTICS , GALIT SHMUELI, PETER C. BRUCE,PETER GEDECK,NITIN R. PATEL

[12] Top 10 algorithms in data mining,Xindong Wu,Jiannong Cao

[13] https://www.cdc.gov/diabetes/basics/diabetes.html

[14] https://www.kaggle.com/datasets/cdc/national-health-and-nutrition-examination-survey

[15] https://www.techtarget.com/searchbusinessanalytics/definition/data-visualizatio

[16] https://www.datacamp.com/tutorial/adaboost-classifier-python

[17] https://scikit-learn.org/stable/modules/svm.html

[18] https://machinelearningmastery.com/confusion-matrix-machine-learning/

[19] https://www.investopedia.com/terms/m/mlr.asp