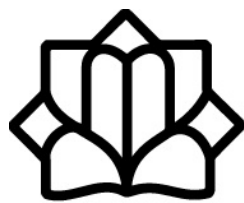




صلى الله عليه وسلم



دانشگاه کاشان

دانشکده مهندسی برق و کامپیوتر

## سامانه پیش‌بینی ابتلا به دیابت (مبتنی بر الگوریتم‌های یادگیری ماشین)

نام و نام خانوادگی دانشجو: قوام الدین سلیمانی

استاد راهنما: سرکار خانم دکتر فرشته دهقانی

زمستان ۱۴۰۱

## تقدیر و تشکر

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

# فهرست مطالب

آ	تقدیر و تشکر
سه	فهرست تصاویر
۴	خلاصه
۵	۱ مقدمه
۵	۱-۱ تاریخچه
۵	۲-۱ علل ابتلا
۶	۱-۲-۱ فاکتور های محیطی
۶	۲-۲-۱ عوامل ژنتیکی و غیر محیطی
۶	۳-۱ راهکار های پیش گیری و استعداد سنجی
۶	۳-۱-۱ مبتلایان به پیش دیابت
۶	۳-۳-۱ سایر افراد جامعه
۷	۴-۱ انگیزه و اهداف انجام این پژوهش
۷	۵-۱ موارد انجام شده
۸	۲ ادبیات پژوهش
۸	۱-۲ مقدمه
۸	۲-۲ الگوریتم های طبقه بندی
۸	۲-۲-۱ ارگرسون لجستیک
۱۲	۲-۲-۲ درخت تصمیم
۱۲	۲-۲-۳ جنگل درختان تصادفی
۱۲	۴-۲-۲ AdaBoost
۱۳	۵-۲-۲ Naive Bayes
۱۳	۳-۲ داده ها و مصور سازی
۱۴	۳-۲-۱ داده های گم شده
۱۴	۳-۳-۲ افراز داده ها
۱۴	۴-۲ استاندارد سازی داده ها
۱۴	۴-۲-۱ امتغیر های ساختگی (دودویی)
۱۴	۴-۲-۲ مقیاس بندی
۱۴	۳-۴-۲ Cross validation
۱۴	۴-۴-۲ مصور سازی
۱۵	۴-۵ انواع نمودار ها و داده های اجمالی
۱۶	۵-۲ نتیجه گیری
۱۷	۳ کارهای پیشین



# فهرست تصاویر

۹	۱-۲ نمودار پراکندگی
۹	۲-۲ نمودار رگرسیون خطی چندگانه در فضا
۱۰	۳-۲ مقایسه رگرسیون لجستیک و خطی
۱۱	۴-۲ نمودار تابع لجستیک
۱۲	۵-۲ مدل Logistic Regression
۱۳	۶-۲ الگوریتم AdaBoost
۱۵	۷-۲ Pairplot

## خلاصه (چکیده)

دیابت یک بیماری مزمن است که فرد مبتلا، قند خون بالاتر از حد مجاز را دارا می‌باشد و این مسئله موجب عوارض و مشکلات جدی در سلامت وی (از جمله برخی نارسایی‌ها، سکته‌ها، آسیب و از کار افتادن اندام‌ها) 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.

روش‌هایی که بتواند به ما کمک کند تا با دقت مناسبی بتوانیم ابتلای افراد مختلف به بیماری را در آینده را پیش‌بینی کنیم، بسیار حائز اهمیت هستند. یادگیری ماشین و داده کاوی با استفاده از داده‌های مختلف که در گذشته جمعه‌آوری شده‌اند 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 اندازه‌گیری کنیم.

داده‌هایی که برای آموزش مدل‌هایمان استفاده کرده‌ایم، از مجموعه دیتاست بیماران هندی (NHANES) که در سال ۲۰۰۹ تا ۲۰۱۲ جمع‌آوری شده بودند، می‌باشد.

یافته‌ها و نتایج:

...



# فصل ۱

## مقدمه

### ۱-۱ تاریخچه

دیابت چیست؟

دیابت یک بیماری مزمن است که 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. هنگامی که دیابت به درستی مدیریت نشود، می‌تواند منجر به عوارض جدی سلامتی مانند بیماری قلبی، انواع سکته مغزی و قلبی، نارسایی کلیه، کوری و آسیب عصبی شود. [۱۲] آمار ابتلا و مرگ و میر بسیار بالایی از این بیماری در جهان وجود دارد و متأسفانه 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 عده زیادی از افراد هم به پیش دیابت مبتلا هستند. [۶] در پیش دیابت، سطح قند خون بالاتر از حد طبیعی است، اما به اندازه کافی برای تشخیص دیابت بالا نیست. پیش دیابت خطر ابتلا به دیابت، بیماری قلبی و سکته را افزایش می‌دهد. [۵] اگر پیش دیابت در افراد وجود داشته باشد، یک برنامه برای تغییر سبک زندگی، می‌تواند به افراد در جلوگیری از این بیماری کمک کند. [۱۲]

این بیماری سه نوع دارد [۵]:

۱. دیابت نوع اول که معروف به دیابت جوانی است چون افراد با سن کمتر از ۳۰ سال معمولاً 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.

### ۲-۱ علل ابتلا

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

## ۱-۲-۱ فاکتور های محیطی

مطابق تحقیقات و بررسی های انجام شده از سال ها پیش تا کنون، عوامل سبک زندگی چون رژیم غذایی نامناسب، عدم فعالیت بدنی و اضافه وزن (مخصوصا میزان توده بدنی) می تواند خطر ابتلا به دیابت را افزایش دهد. [۱۲] همچنین وجود بیماری های زمینه ای مثلا در پانکراس بین افراد می تواند در مبتلا شدن به این بیماری موثر باشد که بنا به تعریف دیابت نوع یک، این عامل مربوط به همین نوع می شود. [۵]

## ۲-۲-۱ عوامل ژنتیکی و غیر محیطی

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 می تواند در اثر بیماری های خانوادگی و ارثی دیگری در افراد وجود داشته باشد. مثل برخی ویروس ها، وجود کلسترول، چربی و فشار خون و... [۳] [۴] [۱۲]

## ۳-۱ راهکار های پیش گیری و استعداد سنجی

مطابق توصیه متخصصین [۳] 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، می توانند از ابتلا به این بیماری پیش گیری کنند.

## ۱-۳-۱ مبتلایان به پیش دیابت

مطابق توصیه پزشکان، در افرادی که به پیش دیابت 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 می توان به موارد ذیل اشاره کرد [۴]:

- حفظ رژیم غذایی غنی از فیبر مثل انواع میوه ها و سبزیجات و کاهش مصرف غذا های شور، چرب و شیرین
- ورزش منظم
- استفاده از برخی دارو ها مطابق تجویز پزشک

## ۲-۳-۱ سایر افراد جامعه

مطابق آمار ها، سالانه بخش دیگری از افراد جامعه که از دسته قبلی سوا بوده اند، به بیماری دیابت مبتلا می شوند. [۳] در این جا با تحلیل برخی فاکتور های سلامتی می توان پیش بینی کرد که آیا این افراد ممکن است با ادامه 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 بررسی های مختلفی انجام شده تا بتوانیم با اندازه گیری برخی فاکتور های کمی و کیفی در افراد، مسئله استعداد در ابتلا به این بیماری را در آنها بررسی کنیم. چالشی که در این زمینه وجود دارد این است که بسیاری از داده های غیر خطی و غیر استاندارد پزشکی با ارتباطات و ساختار های پیچیده وجود دارند که این بررسی ها را دشوار می سازد. [۶]

## ۴-۱ انگیزه و اهداف انجام این پژوهش

در راستای همه موارد مطرح شده در بخش های قبلی ، بر آن شدیم تا تحقیق کنیم با توجه به امکانات و امروزی و دسترسی به مقالات و منابع گوناگون و همچنین توسعه ی ابزار های مبتنی بر یادگیری ماشین و هوش مصنوعی ، درصدد یافتن بهترین راهکارها برای نجات جان انسان های بیشتر با در نظر گیری مناسب ترین الگوریتم ها باشیم .

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

## ۵-۱ موارد انجام شده

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

پس از یافتن مجموعه دادگان نمونه که مربوط به اطلاعات بیماران آمریکایی در سال های ۲۰۰۹ تا ۲۰۱۲ بوده است، با به کار گیری کتابخانه های مختلف پایتون از جمله `Pandas` ، `scikit-learn` ، `numpy` ، `seaborn` و ... عملیات های گوناگونی بر روی داده ها انجام دادیم . 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 فلان با فلان درصد بهترین عملکرد را بین باقی الگوریتم ها در این پیش بینی، شناخته شد .

## فصل ۲

### ادبیات پژوهش

#### ۱-۲ مقدمه

روش‌های نظارت شده ای مانند طبقه بندی و تخمین تلاش می کنند تا رابطه ای میان صفات خاصه و ورودی (که گاه متغیرهای مستقل نامیده می شوند) را با یک یا چند صفت خاصه هدف (که گاه متغیر وابسته نامیده می شود) کشف کنند. در نهایت این رابطه با یک ساختار به عنوان مدل نمایش داده می شود. [۱]

ادامه دارد...

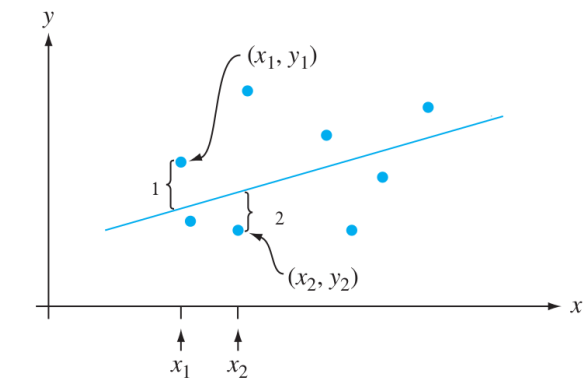
#### ۲-۲ الگوریتم های طبقه بندی

##### ۱-۲-۲ رگرسیون لجستیک<sup>۱</sup>

خوب است در ابتدا مروری بر رگرسیون خطی<sup>۲</sup> داشته باشیم تا در ادامه بتوانیم رگرسیون لجستیک را بهتر درک کنیم.

##### رگرسیون خطی

در بسیاری از بررسی های آماری، لازم است یک متغیر وابسته را از روی یک یا چند متغیر مستقل پیش بینی کنیم که اصطلاحاً به آن رگرسیون یا برگشت می گوئیم. [۲] برای مثلاً میزان ساعت مطالعه یک متغیر مستقل است و نمره اخذ شده در درسی متغیری وابسته است و بین این دو رابطه وجود دارد. سپس نمونه ای 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 می دهیم. [۲] سپس آنها را مثل یک نمودار در صفحه مختصات به یکدیگر متصل کرده که به آن نمودار پراکندگی گوئیم. [۲]



شکل ۲-۱: نمودار پراکندگی

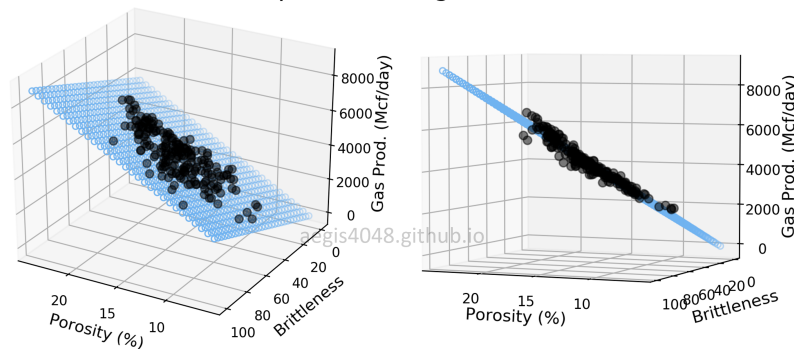
حالا می‌توان خطی را در این صفحه مختصات در نظر گرفت که تا حد زیادی منطبق بر نقاط باشد که در واقع یک نمودار پیش‌بینی کننده  $Y$  بر مبنای  $X$  است که به آن معادله رگرسیون  $Y$  بر روی  $X$  گویند. [۲] حالا رابطه این نقاط و منحنی را با  $\mu_{Y|x} = E(Y|x) = \alpha + \beta x$  مشخص می‌کنیم و  $\alpha$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

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

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

که در این روابط  $\bar{x}$  و  $\bar{y}$  میانگین  $X$  و  $Y$  هستند.

3D multiple linear regression model



شکل ۲-۲: نمودار رگرسیون خطی چندگانه در فضا

با تعمیم این روابط و اصول بیان شده می‌توان حالتی را در نظر گرفت که چندین متغیر مستقل داریم

(مثلاً ۲ تا) که این مدل به رگرسیون خطی چندگانه معروف است و در آنجا نمودار ما حالت فضایی پیدا خواهد کرد و با رابطه زیر می‌توانیم آن را بیان کنیم. [۱] [۲]

$$y_i = \beta_0 1 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, \dots, n \quad (۲-۳)$$

در این رابطه،  $p$  ابعاد ما می‌باشد.

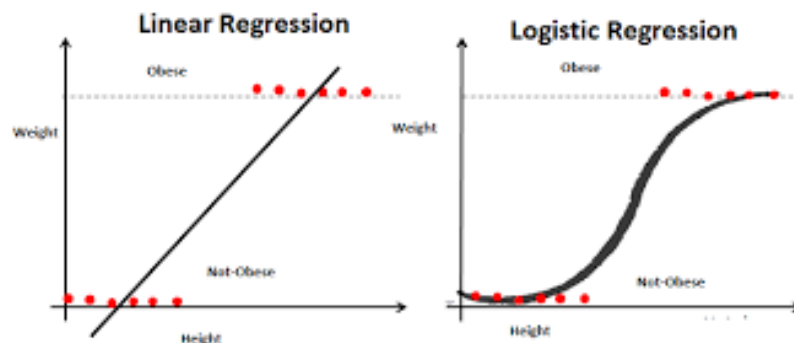
در کل می‌توانیم بگوییم روش‌های رگرسیون زمانی مناسب است که مقادیر مستقل در مجموعه داده‌ها به کلاس‌هایمان (به بیان دیگر طبقه‌بندی‌هایمان) وابستگی داشته باشند. ضریب همبستگی خطی که با رابطه فلان می‌شود میزان این وابستگی را برای ما نشان می‌دهد. [۱]

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

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]

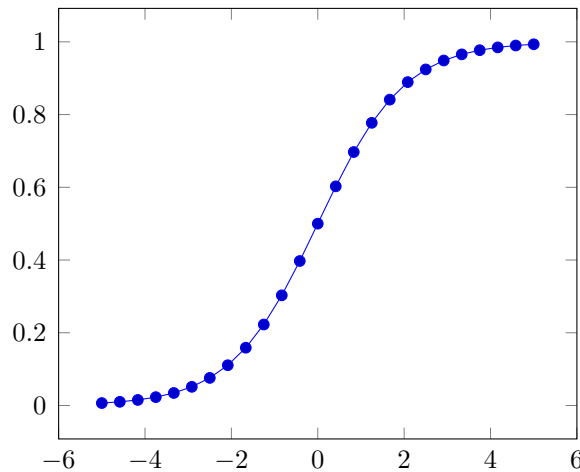
**رگرسیون لجستیک** حالا که با مفاهیم ابتدایی رگرسیون آشنایی پیدا کردیم به بیان نوع دیگری از آن به نام رگرسیون لجستیک می‌پردازیم.

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. مثلاً کار ما، فاکتور BMI در افراد یک متغیر مستقل است و بر ابتلا به دیابت در افراد موثر بوده؛ همچنین در اینجا متغیر وابسته ما، همان دیابت گرفتن یا نگرفتن فرد می‌باشد که با ۰ و ۱ آن را در نظر می‌گیریم. پس این مدل برای مواردی استفاده می‌شود که حالت کلاس‌بندی برای متغیرهایمان داریم. [۸] [۹] ضمناً در حالت کلاس‌بندی، 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 در نظر گرفته می‌شود و امکان نمایش کلاس‌های گوناگون 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 نگرفته‌اند.



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 طبقه‌بندی صحیحی برای ما انجام دهد. [۹] نهایتاً چاره



شکل ۲-۴: نمودار تابع لجستیک

این است که از یک نمودار منحنی شکل برای فشرده کردن نتایج بین ° تا ۱ استفاده کنیم [۹]: رابطه ۲-۵ و شکل ۲-۴

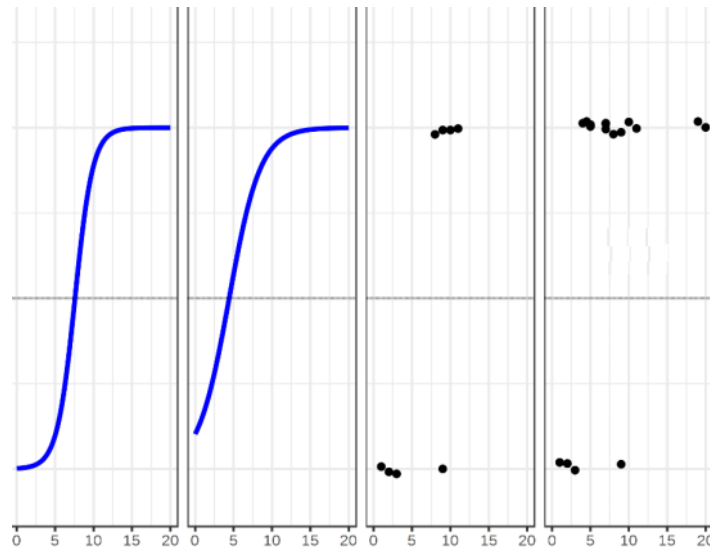
$$\text{logistic}(x) = \frac{1}{1 + \exp(-x)} \quad (۲-۵)$$

سپس اگر سمت راست رابطه ۲-۳ را در جای X در معادلهء بالایی قرار دهیم ۳، فقط مقادیر ° 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 این بخش (۲-۳) و تعمیم خواص آماری (واریانس)، می‌دانیم وزن ویژگی هایمان عاملی اثر گذار تعیین مرز جدا سازی در نمودار ما خواهد بود. لذا نهایتاً به رابطهء زیر می‌رسیم که بر اساس احتمال قرار گیری هر نمونه در هر کلاس برای ما کلاس بندی را انجام می‌دهد [۸] [۹]:

$$g(x) = \ln\left(\frac{p(x)}{1 - P(x)}\right) = \frac{\frac{e^{b_0 + b_1 x}}{1 + e^{b_0 + b_1 x}}}{1 - \frac{e^{b_0 + b_1 x}}{1 + e^{b_0 + b_1 x}}} = \ln(e^{b_0 + b_1 x}) = b_0 + b_1 x \quad (۲-۶)$$

در رابطهء 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 در صورتی که مثلاً با احتمال بالای ۵۰ درصد در طبقه ۱ قرار می‌گیرد، این کار صورت می‌پذیرد و در غیر این صورت به کلاس ° متعلق است. (می‌دانیم که احتمال دقیقاً ۵۰ درصد روی نقطهء (°,°) قرار دارد.) [۸] [۹] به بیان دیگر می‌توانیم بگوییم logit، پاسخ (Y) ترکیب خطی پیش بینی کننده ها یا همان (X) است. [۶]



شکل ۲-۵: مدل Logistic Regression

## ۲-۲-۲ درخت تصمیم

## ۲-۲-۳ جنگل درختان تصادفی

این الگوریتم از ساختار درختان تصمیم استفاده می کند. ساختار درختان تصمیم...

حال می توانیم بگوییم این الگوریتم، با ایجاد چندین درخت تصمیم مختلف از داده هایمان برای ما تصمیم گیری را انجام می دهد. ممکن است پرسید که خب، آیا مثلاً ۴ درخت تصمیم مختلف نتایج یکسانی دارند؟ با توجه به ساختار درختان تصمیم جواب قطعاً خیر است. پس در اینجا از جواب ها رای گیری میشود. یعنی در نمونه ای ۳ درخت حکم می کنند که فرد به دیابت مبتلا خواهد شد و ۱ درخت حکم می کند که این فرد به دیابت مبتلا نخواهد شد. لذا اکثریت آراء بر مبتلا شدن فرد اتفاق نظر دارند. پس نتیجه آن پیش بینی، ابتلا شدن فرد است.

ادامه دارد...

## ۴-۲-۲ AdaBoost

این الگوریتم یک فرض ساده در نظر دارد و آن یادگیری گروهی است. [۱۱] فرض کنید کسانی که در مجموعه داده های آموزشی دچار اشتباه در تشخیص ابتلا به بیماری دیابت شدند، از مجموعه بقیه داده های آموزشی جدا می شوند و در یک طبقه بندی جدید مجدداً مورد ارزیابی قرار می گیرند؛ همانطور که در واقعیت ممکن است هر پزشکی معیار های مختلفی را برای تشخیص بیماری مراجعه کننده اش داشته باشد و هر پزشکی نمی تواند همه ی افراد را درست تشخیص دهد پس اگر یک تیم پزشکی داشته باشیم نظرات جمعی می توانند بیماران بیشتری را به درستی شناسایی کنند. [۱۰]

این روند جداسازی داده های آموزشی و امتیاز دهی در تشخیص توسط طبقه بندی کننده های مختلف درون درختان تصمیم ضعیف، مکرراً تکرار می شود تا در نهایت طبقه بندی های متعددی داشته باشیم که هر کدام بر اساس میزان سرآمدی در تست های مختلف، امتیازات مختلفی دریافت کنند. [۱۱] بعد از ساخته شدن مدل، حالا می توانیم به نسبت امتیاز هر طبقه بندی کننده داده های دریافتی را به صورت شانس بین هر کدام تقسیم کنیم تا پیش بینی صورت گیرد و این مسئله موجب تقویت سنجش می گردد.

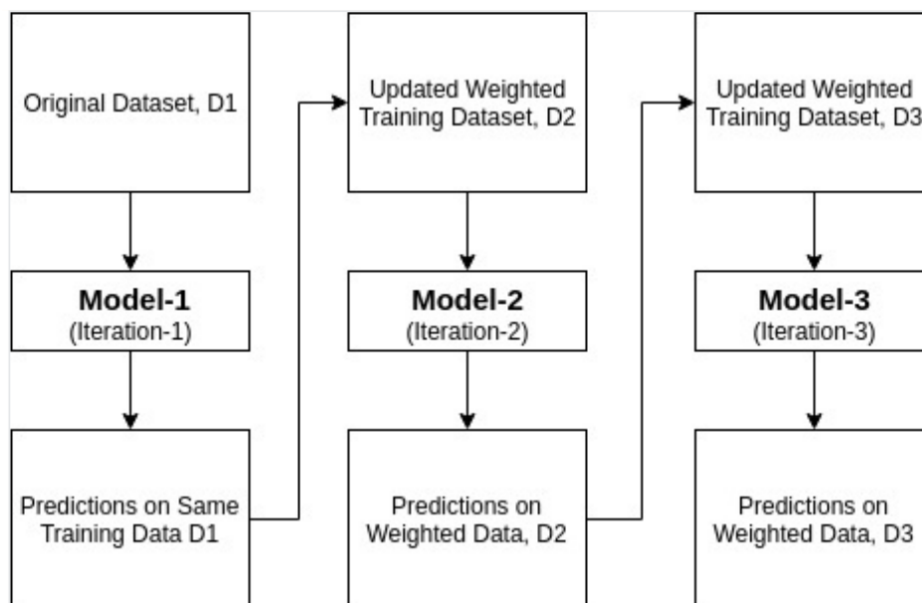


به همین دلیل به آن الگوریتم Adaptive Boosting یعنی تقویت کننده تطبیقی گویند. [۱۰] [۱۱]  
اگر از دید ریاضی الگوریتم را بررسی کنیم، برای محاسبه میزان خطای مدل  $M_i$  وزن هر یک از  
تاپل های  $D_i$  را که  $M_i$  اشتباه طبقه بندی کرد، جمع می کنیم. [۱۰]

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

حالا اگر طبقه بندی کننده  $X_j$  اشتباه کند، میزان ارور برای آن ۱ اندازه گیری می شود و در غیر این  
صورت ۰ است. اگر یک طبقه بندی کننده آن قدر ضعیف باشد که خطایش از ۰.۵ بیشتر شود آن را  
دیگر در نظر نمی گیریم. [۱۰] سپس یک مجموعه داده جدید به نام  $D_j$  تولید می کنیم و از آن یک  $M_i$   
جدید استخراج می کنیم و دوباره این روند را ادامه می دهیم. 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 محاسبه  
وزن هر طبقه بندی استفاده می شود. [۱۰]

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



شکل ۲-۶: الگوریتم AdaBoost [۱۵]: در اینجا مراحل ذکر شده در متن مکررا تکرار می شود.

## ۵-۲-۲ Naive Bayes

## ۳-۲ داده ها 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<sup>۴</sup> تدوین شده است.

<sup>۴</sup>US Centers for Disease Control and Prevention

برنامه NHANES در اوایل دهه ۱۹۶۰ آغاز شد و به صورت مجموعه ای از نظرسنجی ها با تمرکز بر گروه های 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 گیری های سلامت و تغذیه برای رفع نیازهای نوظهور دارد. [۱۳]

مصاحبه NHANES شامل سوالات جمعیت شناختی، اجتماعی-اقتصادی، رژیم غذایی و سلامتی است. جزء معاینه شامل اندازه گیری های پزشکی، دندانپزشکی و فیزیولوژیکی و همچنین تست های آزمایشگاهی است که توسط پرسنل پزشکی بسیار آموزش دیده انجام می شود. [۱۳]

ادامه دارد....

### ۲-۳-۱ داده های گم شده

داده های گم شده چیست؟

راه حل های داده های گم شده

- حذف ردیف های حامل داده های گم شده
  - جایگزینی با متوسط های آماری
- ادامه دارد....

### ۲-۳-۲ افراز داده ها

بخش کردن داده ها یا افراز<sup>۵</sup>

### ۲-۴ استاندارد سازی داده ها

برای ادامهء مراحل، لازم است کارهای بیشتری را روی داده های خود انجام دهیم. از جمله ایجاد متغیر های ساختگی (دو دویی) و مقایس بندی

#### ۲-۴-۱ متغیر های ساختگی (دودویی)

#### ۲-۴-۲ مقیاس بندی

#### ۲-۴-۳ Cross validation

#### ۲-۴-۴ مصور سازی

**تعریف:** به طور ساده می توانیم بگوییم زمانی که داده هایمان را به صورت انواع نمودار ها، نقشه ها و شکل های مختلف بصری دریاوریم تا نتیجه گیری و تحلیل آن ها توسط مغز جهت شناسایی الگو ها و نقاط پرت در داده آسان تر شود، این کار انجام می گیرد. [۱۴]

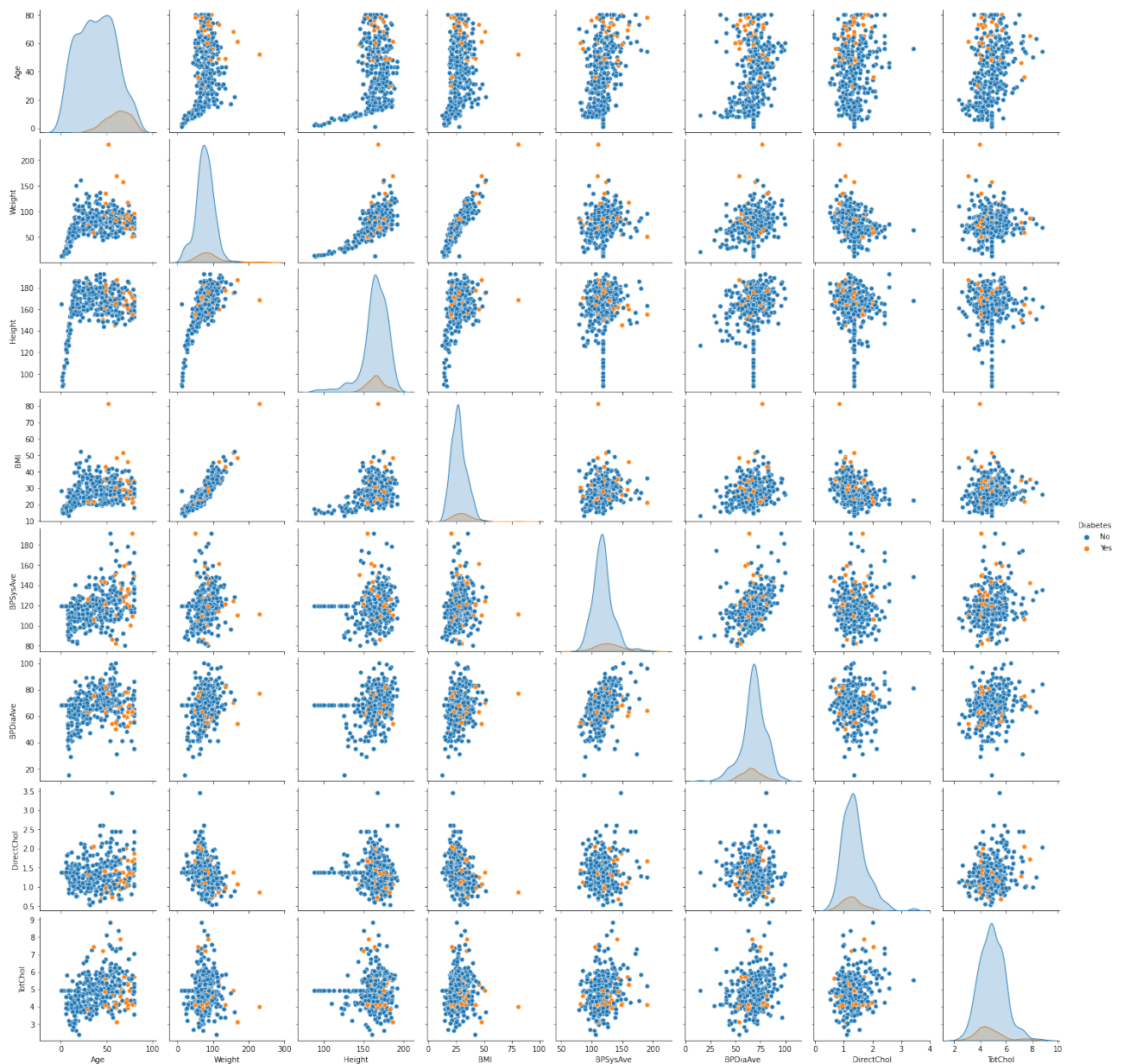
**اهمیت:** یک تصویر هزاران برابر بیشتر از کلمات ارزش دارد.

جملهء فوق، به نوعی اهمیت مصورسازی را برای ما نمایان می کند. درواقع ما با انجام این کار، درک بهتر و سریع تری نسبت به داده های عظیم خود خواهیم داشت. قابل ذکر است که Dataset مورد استفاده ما در اینجا، حدود ۱۰۰۰۰ رکورد را در خود جای داده است.

ضمناً در کنار این موارد، با مصور سازی داده ها دیگر نیازی به توضیحات اضافه نخواهیم داشت و بسیاری از افراد عادی قادر به درک موضوع مطرح شده خواهد بود. [۱۴]

## ۵-۴-۲ انواع نمودار ها و داده های اجمالی

### Pairplot



شکل ۲-۷: Pairplot

در شکل ۲-۷ نمودار Pairplot را مشاهده میکنیم.

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

## ۲-۵ نتیجه گیری

## فصل ۳

### کارهای پیشین

#### ۱-۳ مقدمه

در انجام پژوهش های گوناگون بررسی و مقایسه روش های مختلف در روند توسعه و تحقیق مجدد ، نکته بسیار مهمی است و می تواند اشکالات کار های پیشین را برطرف نمود و در زمینه موارد به روز آن ها را به کار گرفت. همچنین ایده ها و نکاتی در هر مقاله ذکر شده است که می تواند رهنمود های مفید برای کار های آتی تلقی شوند.

#### ۲-۳ مقاله ۱

مقاله [۷] در سال ۲۰۱۸ با استفاده از داده های بیماران هندی پیما<sup>۱</sup> 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 در نظر گرفته شده است.

نکته قابل ملاحظه در این مقاله این است که از الگوریتم SVM هم برای پیش بینی استفاده شده است که معمولاً برای دادگان با مقادیر زیاد روش مناسبی است. [۱۶]

#### ۳-۳ مقاله ۲

در مقالهء اشاره شده [۶] در سال ۲۰۲۰ روش های مختلفی برای پیش بینی ابتلا به دیابت انجام شده و ترکیب جنگل درختان تصادفی و رگرسیون لجستیک در اعتباری سنجی متقابل K-Fold "مکرر" با مقدار  $K=10$  بهترین نتیجه طبقه بندی را با دقت ۹۴٪ حاصل کرده است. در این مقاله روش های بیز ساده و AdaBoost هم مورد استفاده قرار گرفتند که نتایج آن ها داری دقت مناسبی نبوده است.

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

### ۳-۴ نتیجه گیری

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

## فصل ۴

### روش ها و نتایج

۴-۱ مقدمه

۴-۲ روش پیشنهادی

۴-۲-۱ میزان خطای ACC

confusion\_matrix

۴-۳ نتیجه گیری

## فصل ۵

### جمع‌بندی و کارهای آتی

#### ۵-۱ جمع‌بندی

#### ۵-۲ کارهای آتی

در مورد روش‌های توسعه این سیستم می‌توانیم به مواردی چون سیستم‌های یادگیری ماشین آنلاین اشاره کنیم که دادگان همواره با اطلاعات جدید به روزرسانی می‌شوند و در بازه‌های مختلف توسط ناظر سیستم بهترین الگوریتم‌ها بر روی آن‌ها برای کسب بهترین نتایج اعمال می‌گردد. ضمن اینکه می‌توانیم از اینترنت اشیا و امکاناتی از این قبیل استفاده کنیم. به تازگی گجت‌ها و کیت‌های مخصوصی برای گوشی‌های هوشمند طراحی شده‌اند که برای سنجش پارامترهای گوناگون سلامتی مورد استفاده قرار می‌گیرند و با استفاده از گسترش شبکه‌های پرسرعت اینترنت نظیر ۵G به سرعت می‌توانیم حجم عظیمی از داده‌های جدید را دریافت کنیم،

به عنوان یک نمونه ساده و سمبلیک می‌توان از برد الکترونیکی آردوینو که قابلیت نصب گجت‌های مختلف و سنسورهای گوناگون را فراهم می‌آورد استفاده کرد و یک سیستم یادگیری ماشین آنلاین را طراحی کرد که با سرور ما (مثلاً برای آردوینو NodeRed می‌باشد) تبادل دارد،

همچنین می‌توانیم سامانه‌های موازی با این سیستم را نیز راه‌اندازی کرد مثل یک سایت پیش‌بینی‌کننده احتمال ابتلا به بیماری دیابت در افراد که هر شخصی با وارد کردن پارامترهای خودش می‌تواند نسبت به وضعیت سلامتی خود در آینده برآوردی داشته باشد.



```
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')
```

## مراجع

- [۱] اسماعیلی. مهدی، مفاهیم و تکنیک های داده کاوی
- [۲] نعمت الهی. نادر، آمار و احتمالات مهندسی
- [۳] What is diabetes?, Aoife M Egan, Sean F Dinneen
- [۴] Epidemiology of diabetes, Nita Gandhi Forouh, Nicholas J Wareham
- [۵] Diabetes Cookbook FOR DUMMIES 3RD EDITION, by Alan L. Rubin, MD with Cait James, MS
- [۶] Classification and prediction of diabetes disease using machine learning paradigm, Md. Maniruzzaman, Md. Jahanur Rahman, Benojir Ahammed and Md. Menhazul Abedin
- [۷] Prediction of Diabetes using Classification Algorithms, Deepti Sisodia, Dilip Singh Sisodia
- [۸] Logistic Regression, Lynne Connelly
- [۹] Interpretable Machine Learning, Christoph Molnar
- [۱۰] Data Mining Concepts and Techniques, Jiawei Han, Micheline Kamber, Jian Pei
- [۱۱] Top 10 algorithms in data mining, Xindong Wu, Jiannong Cao
- [۱۲] <https://www.cdc.gov/diabetes/basics/diabetes.html>
- [۱۳] <https://www.kaggle.com/datasets/cdc/national-health-and-nutrition-examination-survey>
- [۱۴] <https://www.techtarget.com/searchbusinessanalytics/definition/data-visualization>
- [۱۵] <https://www.datacamp.com/tutorial/adaboost-classifier-python>
- [۱۶] <https://scikit-learn.org/stable/modules/svm.html>