
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

زمستان ۱۴۰۱

تقدیر و تشکر

فهرست مطالب

فهرست تصاویر

خلاصه (چکیده)

دیابت یک بیماری مزمن است که فرد مبتلا ، قند خون بالاتر از حد مجاز را دارا می باشد 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 می باشد.لذا در صورت وجود احتمال ابتلا به این بیماری در افراد، می توان با تغییر سبک زندگی و کنترل های پزشکی ، تا حدی از ابتلای به این بیماری در افراد مستعد و محتمل، جلوگیری کرد.

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 داده کاوی با استفاده از داده های مختلف که در گذشته جمعه آوری شده اند می توانند کمک شایانی به ما در این امر داشته باشند. پس از الگوریتم های 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 کرده‌ایم، از مجموعه دیتاست بیماران هندی (NHANES) که در سال ۲۰۰۹ تا ۲۰۱۲ جمع آوری شده بودند ، می‌باشد.

یافته ها و نتایج:

...

فصل ۱

مقدمه

۱.۱ تاریخچه

دیابت چیست ؟

دیابت یک بیماری مزمن است که زمانی رخ می دهد که بدن انسولین کافی تولید نمی کند یا نمی تواند به طور موثر از انسولین تولید شده استفاده کند. انسولین هورمونی است که به تنظیم سطح قند خون کمک می کند. هنگامی که دیابت به درستی مدیریت نشود، می تواند منجر به عوارض جدی سلامتی مانند بیماری قلبی، انواع سکته مغزی و قلبی، نارسایی کلیه، کوری و آسیب عصبی شود. [۲] آمار ابتلا و مرگ و میر بسیار بالایی از این بیماری در جهان وجود دارد و متاسفانه روز به روز این آمار افزایش می یابد.

طبق آمار ها ، از هر ۱۰ نفر که به دیابت مبتلا هستند ، بیش از ۸ نفر آنها از این مسئله آ گاهی ندارند و عدهء زیادی از افراد هم به پیش دیابت مبتلا هستند . 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 می دهد. اگر پیش دیابت در افراد وجود داشته باشد ، یک برنامه برای تغییر سبک زندگی ، می تواند به افراد در جلوگیری از این بیماری کمک کند.

این بیماری سه نوع دارد :

۱. دیابت نوع اول که معروف به دیابت جوانی است چون افراد با سن کمتر از ۳۰ سال معمولا مبتلا می شوند . در این نوع به طور ساده می توانیم 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 سلول ها می باشد.

۳. دیابت نوع سوم دیابت بارداری است که در خانم های باردار به طور موقت اتفاق می افتد.

۲.۱ علل ابتلا

در ابتلا به این بیماری بنا به نوع آن و همچنین شرایط ژنتیکی و محیطی افراد مختلف ، فاکتور های مختلفی مطرح است:

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. 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.

۲.۲.۱ 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 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 ...

۳.۱ راهکار های پیش‌گیری و استعدادسنجی

مطابق توصیهء 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 ،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 :

- 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 دستهء قبلی سوا بوده اند ،به بیماری دیابت مبتلا می شوند. در این جا با تحلیل برخی فاکتور های سلامتی می توان پیش‌بنی کرد که آیا این افراد ممکن است با ادامهء 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 ، مسئلهء 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 غیرخطی و غیراستاندارد پزشکی با ارتباطات و ساختارهای پیچیده وجود دارند که این بررسی ها را دشوار می‌سازد.

لذا در این بررسی ما سعی کردیم از انواع الگوریتم های یادگیری ماشین مثل KNN ، جنگل درختان تصادفی و ... استفاده کنیم تا ببینیم کدام الگوریتم ها ، نتایج دقیق تر و پیش‌بینی های بهتری را برای ما به ارمغان می آورد؟

فصل ۲

داده‌ها و مصورسازی

۱.۲ مقدمه

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[1] تدوین شده است.

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.[۳]

مصاحبه NHANES شامل سوالات جمعیت شناختی، اجتماعی ـاقتصادی، رژیم غذایی و سلامتی است. جزء معاینه شامل اندازه گیری های پزشکی، دندانی و فیزیولوژیکی و همچنین تست های آزمایشگاهی است که توسط پرسنل پزشکی بسیار آموزش دیده انجام می شود.[۳]

ادامه دارد....

۲.۲ مصورسازی

۱.۲.۲ تعریف و اهمیت مصورسازی

تعریف: به طور ساده می توانیم بگوییم زمانی که داده هایمان را به صورت انواع نمودار ها ، نقشه ها و شکل های مختلف بصری دربیاوریم 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. [۴]

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

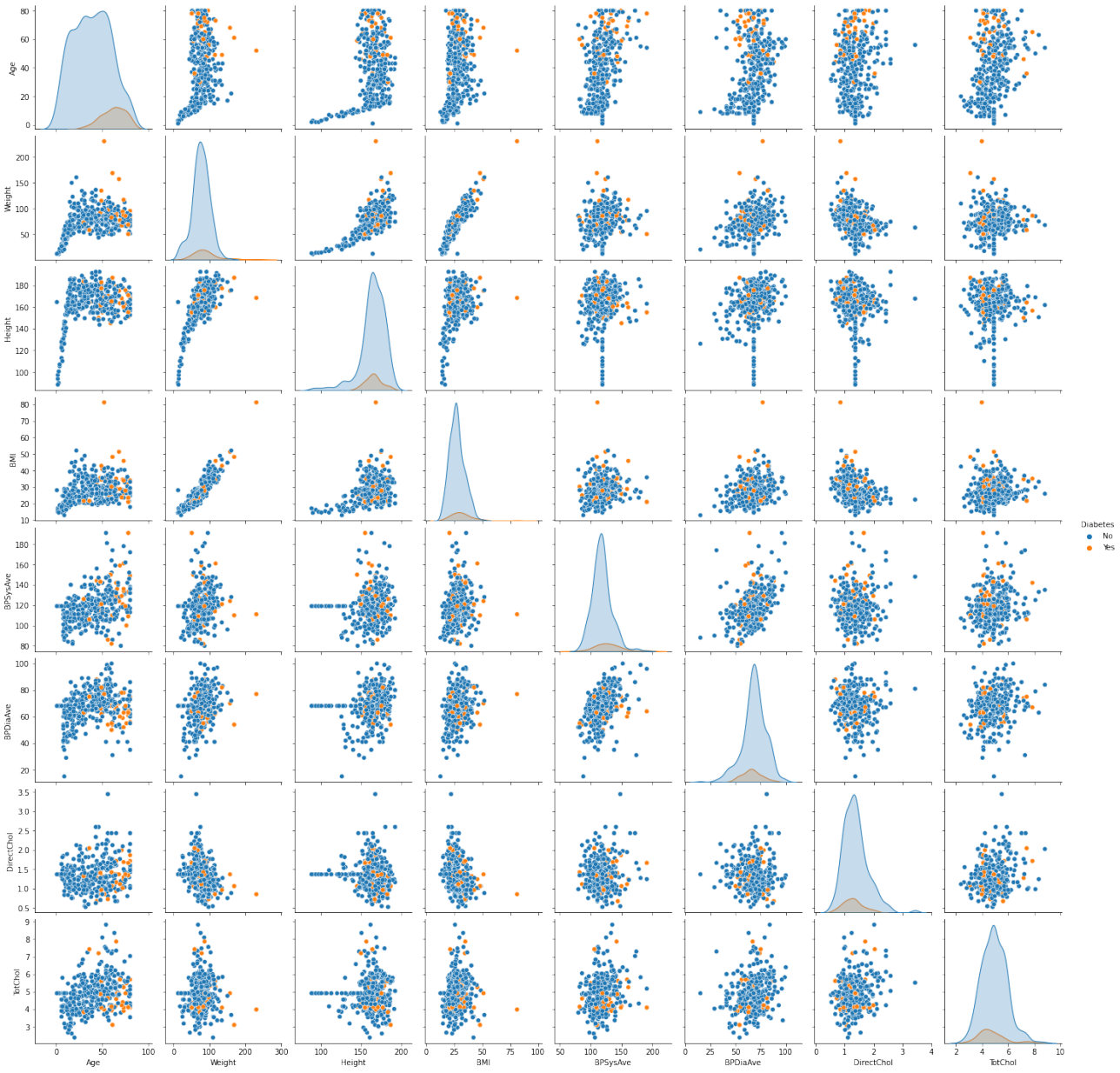
جملهء فوق، به نوعی اهمیت مصورسازی را برای ما نمایان می کند.درواقع ما با انجام این کار ، درک بهتر و سریع تری نسبت به داده های عظیم خود خواهیم داشت.قابل ذکر است که Dataset مورد استفاده ما در اینجا ، حدود ۱۰۰۰۰ رکورد را در خود جای داده است.

[1] US Centers for Disease Control and Prevention

ضمنا 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. [۴]

۳.۲ انواع نمودار ها و داده های اجمالی

۱.۳.۲ Pairplot



شکل ۱.۲: Pairplot

در شکل ۱.۲ نمودار Pairplot را مشاهده میکنیم.

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

ادامه دارد....

۴.۲ داده های گم شده

۱.۴.۲ داده های گم‌شده چیست؟

۲.۴.۲ راه حل های داده های گم شده

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

ادامه دارد....

۳.۴.۲ افراز داده ها

بخش کردن داده ها یا افراز [۲]

[۲] Data Splitting

فصل ۳

معرفی روش ها و الگوریتم ها

۱.۳ مروری بر روش های یادگیری با نظارت و بی نظارت

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

ادامه دارد....

۲.۳ استاندارد سازی داده ها

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

۱.۲.۳ متغییر های ساختگی (دودویی)

۲.۲.۳ مقیاس بندی

۳.۳ الگوریتم های طبقه بندی

۱.۳.۳ رگرسیون لجستیک

خوب است در ابتدا مروری بر رگرسیون خطی داشته باشیم.

۲.۳.۳ 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 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 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 ی آن پیش بینی، ابتلا شدن فرد است.

ادامه دارد...

فصل ۴

نتیجه گیری

۱.۴ مقایسه میزان خطا در الگوریتم های مختلف

confusion_matrix

۲.۴ پیشنهادات

پیشنهاد

پیوست آ

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 بر روی آن ، در فایل خروجی شاهد قرار گرفتن عبارات انگلیسی درون تگ lr خواهیم بود.

```
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

Data Spliting: افراز داده‌ها

مراجع

[۱] اسماعیلی.مهدی ، مفاهیم و تکنیک های داده کاوی

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

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

[۴] https://www.techtarget.com/searchbusinessanalytics/definition/data-visualization