دانشکده مهندسی

گروه مهندسی صنایع

نام درس : تحلیل آماری کاربردی

نام استاد : دکتر علیرضا شادمان

نام پروژه : پیش بینی بیماری دیابت بر اساس روش‌های طبقه بندی

دانشجو :

امیرعلی باقرزاده بیوکی- 9912743386

تیر 1402


بسم الله الرحمن الرحيم

فهرست

1-مقدمـه 1

2-تشریح داده‌ها 1

3-آماده‌سازی اولیه 4

4-پیاده‌سازی روش‌ها و ارزیابی 7

4.1- معرفی روش پیشنهادی - جنگل تصادفی 7

4.2- معرفی روش پیشنهادی – k-fold cross validation 7

5-نتیجه‌گیری و جمع بندی 15

6-منابع 15

پیش بینی بیماری دیابت بر اساس روش‌های طبقه بندی

امیرعلی باقرزاده بیوکی 9912743386 - amiralibghd1881@gmail.com

1-مقدمه

دیابت یکی از بیماری‌هایی است که اگر فردی دچار آن شود تا پایان عمر درگیر آن است. بنابراین پیشگیری از این بیماری می‌تواند بهترین گزینه برای هر فردی باشد تا از دچار شدن به این بیماری جلوگیری کند. روش های یادگیری و طبقه بندی این کار را بر اساس سوابق پزشکی بیمار انجام می‌دهد و می‌تواند پیش‌بنی کند که آیا با این سوابق پزشکی فرد دچار دیابت خواهد شد یا خیر سپس با رژیم غذایی مناسب از این بیماری پیش‌گیری کند.

2-تشریح داده‌ها

این پروژه با استفاده از مجموعه داده‌های به دست آمده از سایت Kaggle[1] تکمیل شده است. این مجموعه داده‌ها حاوی اطلاعات پزشکی و دموگرافیک بیماران است .

از ویژگی‌های(متغیر) مختلفی مانند سن، جنسیت، شاخص توده بدنی (BMI)، فشار خون بالا، بیماری قلبی، سابقه مصرف سیگار، سطح HbA1c و سطح گلوکز خون تشکیل شده است.

متغیر پاسخ همان وضعیت دیابتی بودن است که یک متغیر کیفی است. اگر این متغیر 1 باشد به این معنی است که فرد مورد نظر دیابتی است و اگر 0 باشد به این معنی است که فرد مورد نظر دیابتی نیست.

این داده دارای 100/000 ریکورد است که به دلیل زمان طولانی در اجرای کدهای استفاده شده ما 1204 رکورد از این داده را به صورت تصادفی انتخاب کردیم و مدل‌ها را رو این تعداد داده پیاده سازی کردیم.

این مجموعه داده دارای داده گمشده‌ای نیست.

متغیر سابقه مصرف سیگار دارای 6 دسته است که به ترتیب never ، ever، current ، former، not current و no info نام گذاری شده‌اند. به دلیل اینکه سلامتی انسان از اهمیت ویژه‌ای برخوردار است در داده‌ای که برای این پروژه در نظر گرفتیم رکوردهای مرتبط با no info را حذف کردیم و در داده خود نیاوردیم. همچنین به دلیل سادگی در تحلیل مفاهیم مرتبط با این مجموعه داده، سابقه مصرف سیگار را به دو دسته هرگز و بله تقسیم بندی کردیم.

در ارتباط با تاریخ جمع‌آوری داده اطلاعاتی در دسترس نیست.

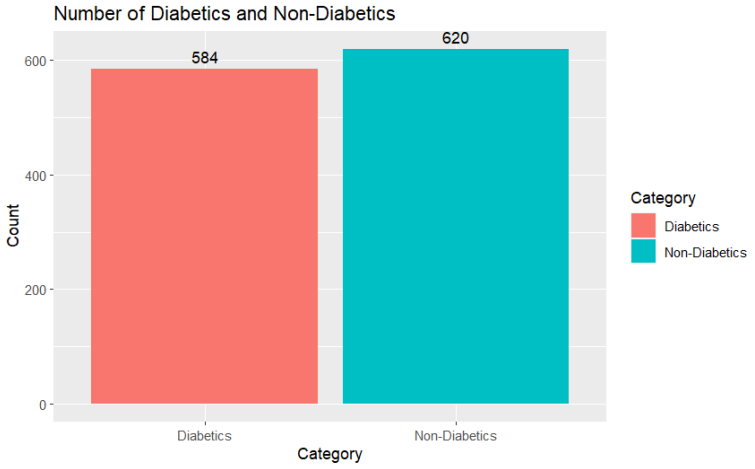
در جدول 1 به شرح کوتاهی از متغیرهای این مجموعه داده می‌پردازیم.

جدول 1- شرح کوتاهی از متغیرها

متغیر	شرح مختصر	نوع متغیر
Gender	جنسیت شامل female(زن) و male(مرد)	factor
Age	سن	numeric
hypertension	دارای فشار خون (1=بله، 0=خیر)	factor
heart_disease	سابقه بیماری قلبی(1=بله، 0=خیر)	factor
smoking_history	سابقه سیگاری بودن(1=بله، 0=هرگز)	factor
Bmi	شاخص bmi: قد^2/وزن	numeric
HbA1c_level	سطح HbA1c (هموگلوبین A1c) معیاری از میانگین سطح قند خون یک فرد در 2 تا 3 ماه گذشته است.	numeric
blood_glucose_level	سطح گلوکز خون	numeric
diabetes	متغیر پاسخ، وضعیت دیابتی بودن فرد(1=بله، 0=خیر)	factor

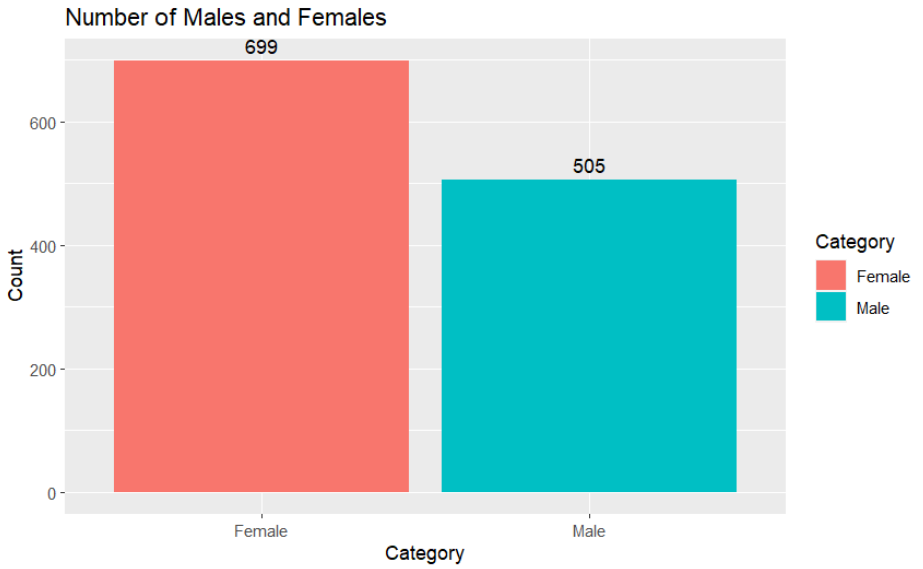
ابتدا با استفاده از نمودار نگاه کلی به مجموعه داده می‌اندازیم و سپس داده‌ها را تقسیم‌بندی می‌کنیم و متغیرها را با یکدیگر بررسی می‌کنیم.

نمودار میله ای متغیر دیابت، در تصویر 1 نشان داده شده است همانطور که ملاحظه می‌شود 584 نفر دیابتی هستند و 620 نفر دیابتی نیستند



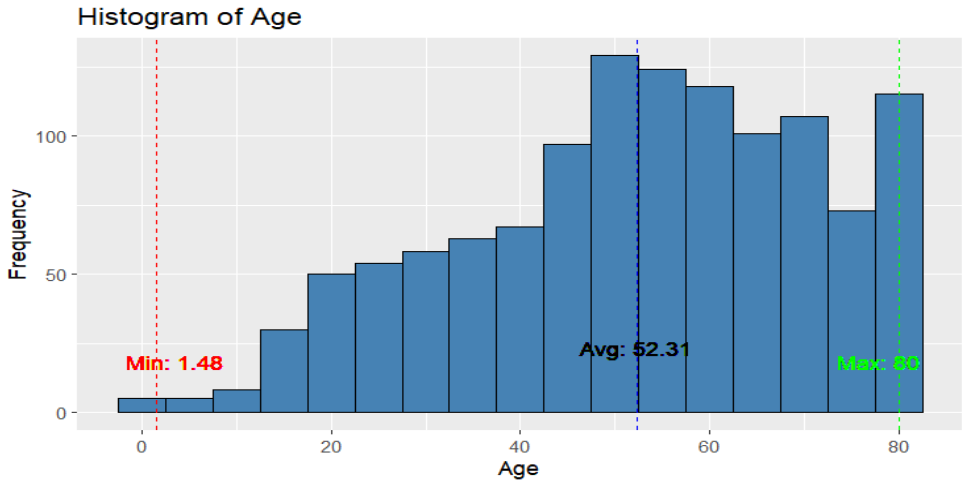
تصویر 1- نمودار میله‌ای متغیر دیابت

نمودار میله ای متغیر کیفی جنسیت، در تصویر 2 نشان داده شده است همانطور که ملاحظه می‌شود 505 نفر مرد هستند و 699 نفر زن هستند.



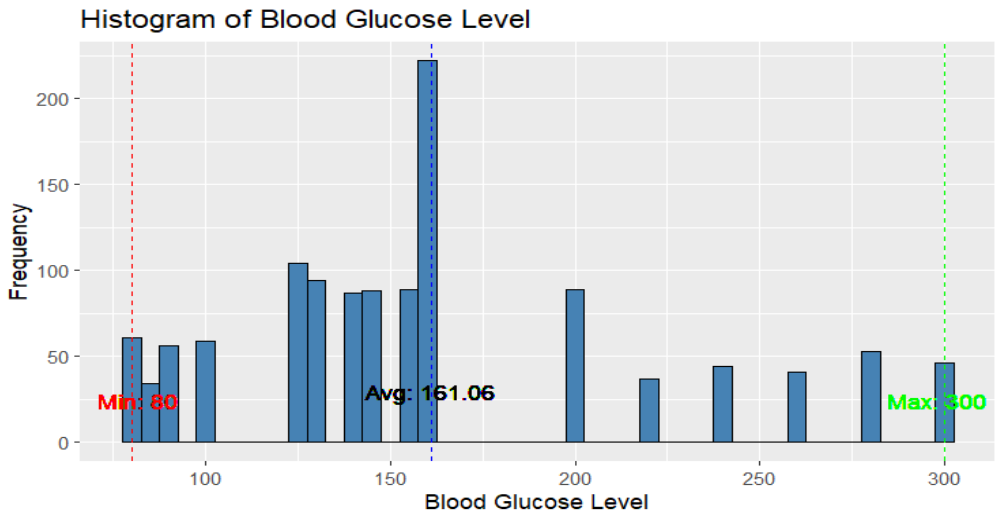
تصویر 2- نمودار میله‌ای متغیر جنسیت

همانطور که در تصویر 3 مشاهده می‌شود در این مجموعه داده کمترین سن 1.48 سال و بیشترین سن 80 سال داست و میانگین سنی 52 سال است.



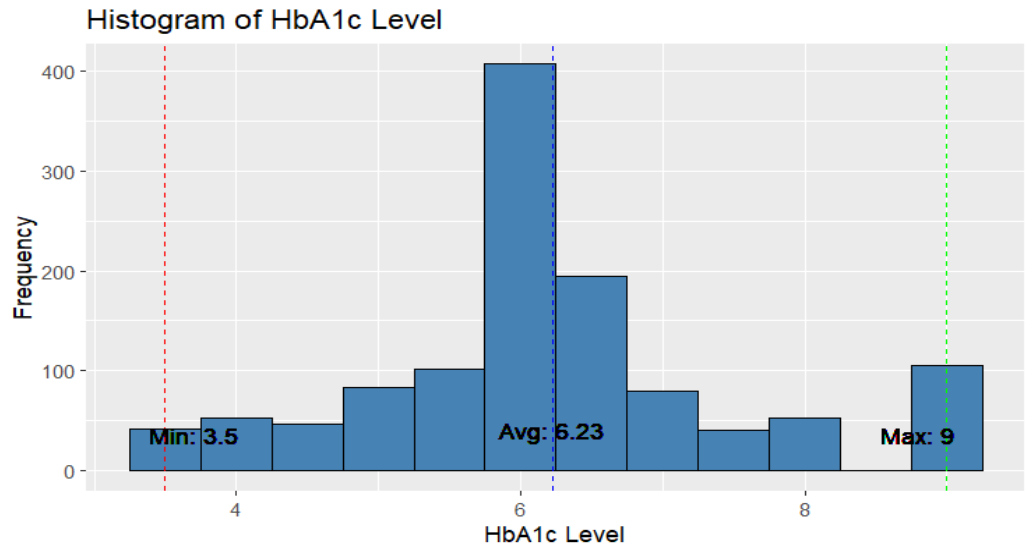
تصویر 3- نمودار هیستوگرام متغیر سن

همانطور که در تصویر 4 مشاهده می‌شود در این مجموعه داده کمترین سطح گلوکز خون 80 و بیشترین سطح گلوکز خون 300 است و میانگین این متغیر 161.06 است.



تصویر 4- نمودار هیستوگرام متغیر سطح گلوکز خون

و آخرین متغیری که مهم است و باید اطلاعاتی در این قسمت درباره آن داده شود سطح HbA1c (هموگلوبین A1c) همانطور که در تصویر 5 مشاهده می‌شود در این مجموعه داده کمترین سطح HbA1c (هموگلوبین A1c) عدد 3.54 است و بیشترین آن عدد 9 و میانگین این متغیر 6.23 است.



تصویر 5- نمودار هیستوگرام متغیر سطح HbA1c (هموگلوبین A1c)

3-آماده‌سازی اولیه داده‌ها

ابتدا متغیر پاسخ که دیابتی بودن است را تبدیل به متغیر فاکتور می‌کنیم. همچنین متغیرهای فشارخون(hypertension)، بیماری قلبی(heart disease) و تاریخچه سیگاری بودن(smoking_history) را به متغیر فاکتور تبدیل می‌کنیم.

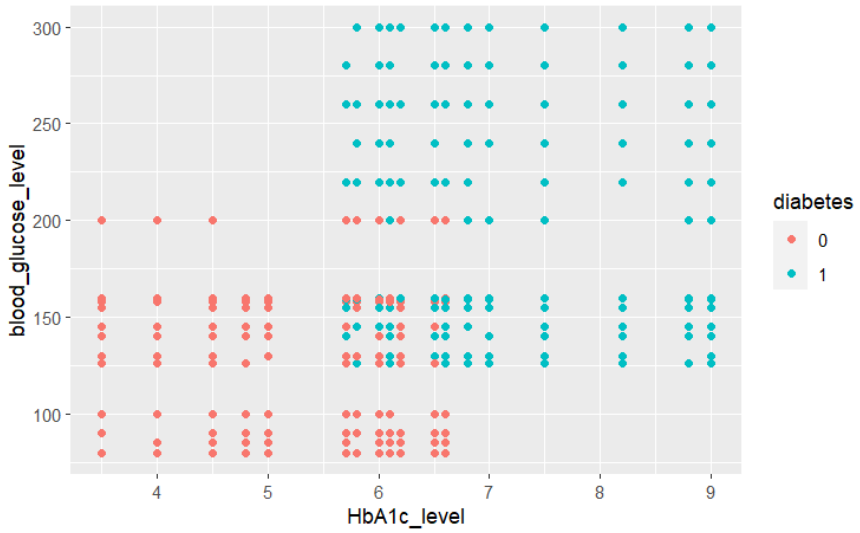
به جهت ارزیابی مدل‌ها و صحت‌سنجی، داده‌ها را به دو مجموعه آموزشی (train) و آزمون (test) تقسیم‌بندی می‌کنیم به طوریکه بهص ورت تصادفی 80 درصد از داده آموزشی و 20 درصد از داده آزمون هستند همچنین مجموعه داده آموزشی را به مجموعه‌های برآورد((estimation) و اعتبار سنجی (validation) تقسیم‌بندی می‌کنیم که به صورت تصادفی 80 درصد از داده برآورد و 20 درصد از داده اعتبارسنجی هستند.

برای پیاده‌سازی مدل KNN متغیرهای سن، bmi ، سطح گلوکز خون و سطح HbA1c را مقیاس‌بندی(scale) کردیم.

حال رابطه بین متغیرها را با نمودارها بررسی می‌کنیم.

ابتدا رابطه بین سطح HbA1c (هموگلوبین A1c) و سطح گلوکز خون را بررسی می‌کنیم.

همانطور که در تصویر 6 مشاهده می‌شود تقریبا از سطح گلوکز خون 200 و بالاتر و سطح HbA1c (هموگلوبین A1c) 7 و بالاتر بیماران ما دیابتی محسوب می‌شوند و این نمودار ارتباط بسیار خوب این دو متغیر با یکدیگر را نشان می‌دهد.



تصویر 6- نمودار رابطه بین سطح گلوکز خون و سطح HbA1c (هموگلوبین A1c)

رابطه بعدی که قصد بررسی آن را داریم رابطه‌ی بین متغیر سن و سطح گلوکز خون است طبق این نمودار که در تصویر 7 مشاهده می‌شود که اکثر بیماران دیابتی ما بالای 40 سال سن دارند و همچنین سطح گلوکز خون آن‌ها بالاتر از 200 است. می‌توان گفت رابطه‌ای میان این دو متغیر وجود دارد اما همچنان رابطه بین سطح گلوکز خون و سطح HbA1c قوی‌تر و بهتر است.

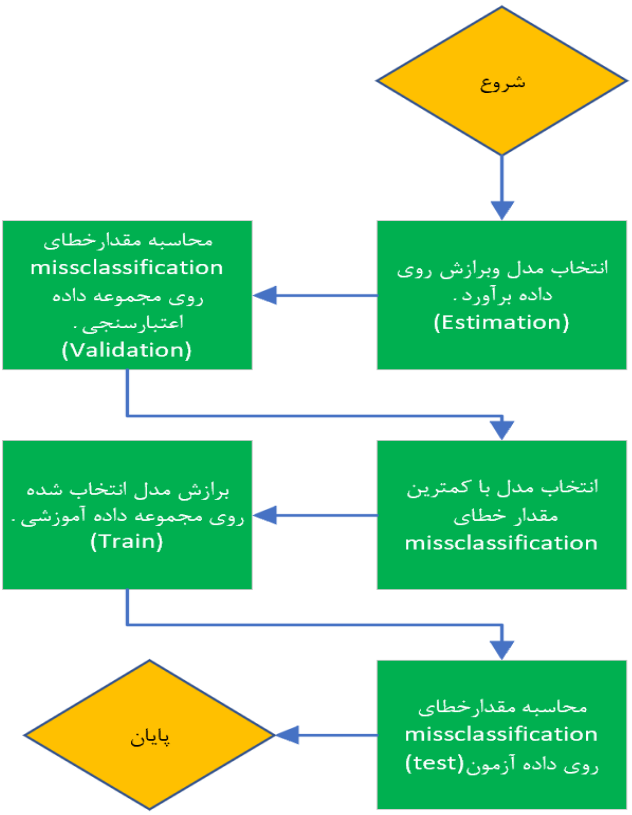
تصویر7- نمودار رابطه بین سطح گلوکز خون و سن

نمودار دیگری که می‌توان رسم کرد و رابطه بین دو متغیر مهم سطح گلوکز خون و سطح HbA1c (هموگلوبین A1c) نشان داد در تصویر 8 مشاهده می‌شود.

تصویر 8- نمودار رابطه بین سطح گلوکز خون و سطح HbA1c (هموگلوبین A1c)

4-پیاده‌سازی روش‌ها و ارزیابی

فرآیند این بخش را در دیاگرام زیر مشاهده می‌کنید.

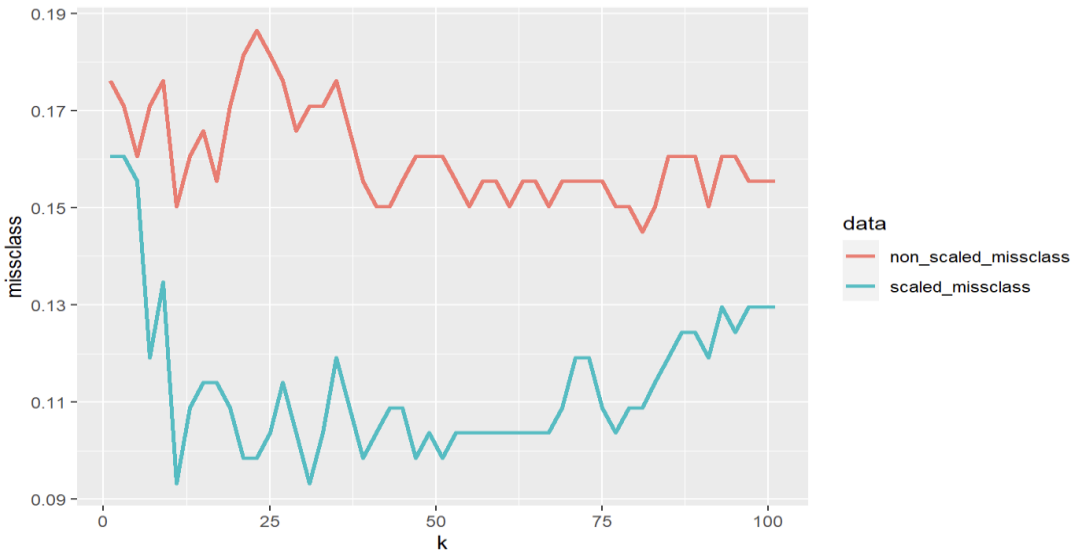


تصویر 9- فرآیند مدلسازی و انتخاب بهترین مدل

حال به بررسی روش‌های مختلف برای پیش‌بینی کلاس‌ها می‌پردازیم.

روشKNN

برای انتخاب بهترین مدل KNN ابتدا دو مدل ایجاد کرده که در یکی از آن‌ها داده‌ها رامقیاس‌بندی کردیم و در مدل دیگر مقیاس‌بندی انجام ندادیم. سپس این مدل‌ها را بروی مجموعه داده برآورد(estimation) برازش دادیم و نرخ خطای طبقه‌بندی نادرست را بر روی مجموعه داده اعتبارسنجی به ازای مقادیر فرد K از 1 تا 101 محاسبه کردیم. جزئیات در جدول 2 برای نمونه آمده است. همچنین روند خطا را نیز می‌توان در تصویر 10 دید.



تصویر 10- روند خطای طبقه‌بندی در مجموعه مقیاس‌بندی شده و غیر مقیاس‌بندی شده

جدول2- خطا با داده مقیاس‌بندی شده وغیرمقیاس‌بندی شده برای مقادیر مختلف k

K	non_scaled_missclass	scaled_missclass
1	0.1761658	0.16062176
3	0.1709845	0.16062176
5	0.1606218	0.15544041
7	0.1709845	0.11917098
9	0.1761658	0.13471503
11	0.1502591	0.09326425

از جدول 2 و تصویر 10 نتیجه می‌گیریم که با مقیاس‌بندی داده خطای مورد نظر کمتر می‌شود. در جدول زیر بهترین خطای طبقه‌بندی در مجموعه اعتبارسنجی ثبت شده است.

جدول3- بهترین مدل knn

مدل	مقدار k	نرخ خطای طبقه‌بندی نادرست داده اعتبارسنجی	نرخ خطای طبقه‌بندی نادرست داده برآورد
knn	11	0.09326425	0.112987

در جدول زیر مقادیر پیش‌بینی در برابر مقادیر واقعی نمایش داده شده است.

جدول4- مقادیر واقعی در برابر مقادیر پیش‌بینی شده در مدل KNN به ازای k = 11

		مقادیر واقعی	
		0	1
مقادیر پیش‌بینی شده	0	97	13
	1	5	78

مجموع مقادیری که اشتباه پیش‌بینی شده‌اند، 18 می‌باشد که نسبت به کل تقریبا 9 درصد خطا موجود است. از طرفی این خطا در افراد بیمار بیشتر است. KNN توانسته افراد سالم را به خوبی پیش‌بینی کند اما 13 نفر از بیماران را نیز سالم پیش‌بینی کرده است.

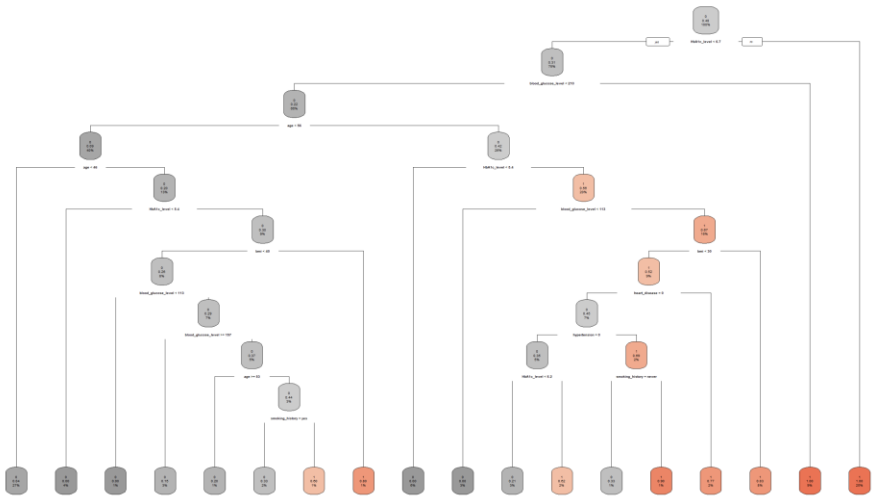
روش درخت تصمیم

برای انتخاب بهترین مدل تصمیم ابتدا مدل‌ها را بر روی مجموعه داده برآورد برازش می‌کنیم و به ازای مقادیر 0.001، 0.01، 0.1 و 1 برای cp و مقادیر 5،10،15،20،25،30 برای minsplit کدهایی را نوشتیم که خلاصه آن در جدول زیر آمده است.

جدول5- cp،minsplit و خطا در درخت تصمیم

خطا	cp	minsplit
0.119171	0.001	15
0.1295337	0.01	5
0.1813472	0.1	5
0.4715026	1	5

همانطور که در جدول 3 مشاهده می‌شود کمترین میزان خطا برابر است به ازای cp=0.001 و minsplit=15 برابر است با 0.119171 .همچنین نمودار درخت تصمیم در شکل زیر نشان داده شده است.



تصویر 10- نمودار درخت تصمیم

بر اساس درخت و همینطور استفاده از variable.importance متوجه می‌شویم که اهمیت متغیرها در این مجموعه داده به ترتیب عبارتند از HbA1c_level، blood_glucose_level، age،bmi، hypertension و heart disease.

در جدول 6 مقادیر پیش‌بینی شده در مقابل مقادیر واقعی توسز درخت را مشاهده می‌کنیم.

جدول6- مقادیر واقعی در برابر مقادیر پیش‌بینی شده در مدل درخت تصمیم

	مقادیر واقعی		
		0	1
مقادیر پیش‌بینی شده	0	89	10
	1	13	81

درجدول بالا نشان داده شد که 13 نفر از افراد سالم بیمار و 10 نفر از افراد دارای دیابت سالم پیش‌بینی شده‌اند.

4.1-معرفی روش پیشنهادی - جنگل تصادفی

مدل‌های رگرسیون مانند درخت تصمیم، در معرض بیش برازش قرار دارند و این مدل برای برطرف کردن نقاط ضعف درخت تصمیم توسعه داده شده است. این روش بر مبنای ترکیب چندین درخت تصمیم تشکیل شده است و به عنوان یک الگوریتم قوی و پرکاربرد در حوزه یادگیری ماشین شناخته می‌شود. حال می‌خواهیم یک مدل با جنگل تصادفی بسازیم. توضیحات بیشتر در مورد جنگل تصادفی و تابع آن در فایل کد موجود است.

جدول6- نرخ‌های خطای طبقه‌بندی نادرست در مدل جنگل تصادفی

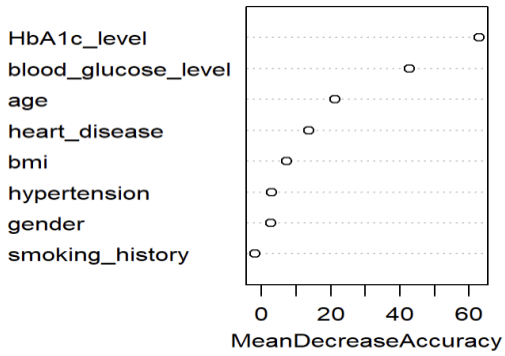
مدل	تفسیر مدل	نرخ خطای طبقه‌بندی نادرست داده اعتبارسنجی	نرخ خطای طبقه‌بندی نادرست داده برآورد
forest	در این مدل متغیر دیابت را در برابر تمام متغیرها بر روی داده برآورد برازش دادیم	0.0880829	0.02467532

همچنین در جدول 7 می‌توانید شاهد مقادیر پیش‌بینی در برابر مقادیر واقعی باشید.

جدول7- مقادیر واقعی در برابر مقادیر پیش‌بینی شده در مدل جنگل تصادفی

	مقادیر واقعی		
		0	1
مقادیر پیش‌بینی شده	0	92	7
	1	10	84

مدل جنگل تصادفی مدل خوبی است. همانطور که مشاهده می‌کنید در جدول 6، خطای طبقه‌بندی را تا 0.02 در مجموعه برآورد و 0.088 در مجموعه اعتبارسنجی کاهش داده است که نشان‌دهنده عملکرد خوب این مدل است و در مجموع 17 مشاهده را نتوانسته که به درستی پیش‌بینی کند. همچنین ما می‌توانیم از آن برای انتخاب متغیر برای مدل استفاده کنیم. در تصویر 11 اهمیت متغیرها ترسیم شده است.



تصویر 11- اهمیت متغیرها

همانطور که مشاهده می‌شود، مهمترین متغیر HbA1c_level است و سپس موارد دیگر می‌باشد. در رابطه با سنجیدن اهمیت ها، در ابتدا خطای MSE مدل ثبت می‌شود. حال مدل را با حذف هر یک از متغیرها می‌سازد و مقدار incRMSE% نشان‌دهنده این است که در نبود هر متغیر چه درصدی خطای MSE افزایش خواهد یافت.

با توجه به اهمیت متغیرها در ادامه مدل log_reg_2 ساخته خواهد شد.

روش رگرسیون لجستیک(logestic regression)

ابتدا با استفاده از تابع multinom مدل خود را روی داده‌های برآورد برازش داده و سپس پیش‌بینی را با داده‌های اعتبارسنجی و برآورد انجام می‌دهیم و در آخر نیز نرخ خطای طبقه‌بندی نادرست را روی داده‌های اعتبارسنجی و برآورد محاسبه می‌کنیم در جدول شماره 8 جزئیات آورده شده است. لازم به ذکر است که یک بار متغیر پاسخ را در برابر همه متغیرها قرار دادیم و یک بار هم در برابر دو متغیر اصلی و اساسی یعنی سطح گلوکز خون و هموگلوبین خون. با توجه به نرخ خطای طبقه‌بندی نادرست مدل log_reg_1 انتخاب می‌شود.

جدول8- نرخ‌های خطای طبقه‌بندی نادرست در مدل رگرسیون لجستیک

مدل	تفسیر مدل	نرخ خطای طبقه‌بندی نادرست داده اعتبارسنجی	نرخ خطای طبقه‌بندی نادرست داده برآورد
log_reg_1	در این مدل متغیر دیابت را در برابر تمام متغیرها بر روی داده برآورد برازش دادیم	0.119171	0.1142857
log_reg_2	در این مدل متغیر دیابت را در برابر متغیر سطح گلوکز خون وسطح هموگلوبین خون بر روی داده برآورد برازش دادیم	0.1865285	0.1532468

در جدول زیر، مقادیر پیش‌بینی در برابر مقادیر واقعی نمایش داده شده است.

جدول9- مقادیر واقعی در برابر مقادیر پیش‌بینی شده در مدل رگرسیون log_reg_1

		مقادیر واقعی	
		0	1
مقادیر پیش‌بینی شده	0	91	12
	1	11	79

جدول10- مقادیر واقعی در برابر مقادیر پیش‌بینی شده در مدل رگرسیون log_reg_2

		مقادیر واقعی	
		0	1
مقادیر پیش‌بینی شده	0	90	26
	1	12	65

با توجه به جدول 2 مشاهده می‌شود که مدل log_reg_1 دقت بالاتری نسبت به مدل دوم دارد. مدل‌های بعدی به این خاطر که دقت نسبتا بالایی نداشته‌اند، مقادیر واقعی آن‌ها در گزارش نیامده است و درصورت نیاز به کد مراجعه شود.

مدل LDA

برای انتخاب بهترین مدل LDA ابتدا مدل را بر روی مجموعه داده برآورد برازش می‌دهیم و سپس خطا را محاسبه می‌کنیم. جزئیات در جدول 11 آمده است.

جدول11- مدل LDA

مدل	تفسیر مدل	نرخ خطای طبقه‌بندی نادرست داده اعتبارسنجی	نرخ خطای طبقه‌بندی نادرست داده برآورد
Lda_1	در این مدل متغیر دیابت را در برابر تمام متغیرها بر روی داده برآورد برازش دادیم	0.1243523	0.1181818

اگر احتمال‌ها را برابر با 0.5 در نظر بگیریم خطاهای مدل ما به صورتی که در جدول 7 آمده است می‌شود.

جدول12- مدل LDA

مدل	تفسیر مدل	نرخ خطای طبقه‌بندی نادرست داده اعتبارسنجی	نرخ خطای طبقه‌بندی نادرست داده برآورد
Lda_2	در این مدل متغیر دیابت را در برابر تمام متغیرها بر روی داده برآورد برازش دادیم	0.119171	0.1181818

مشاهده می‌شود که خطای برآورد یکسان است اما خطای اعتبارسنجی متفاوت است و برای مدل با احتمالای برابر کمتر از حالت عادی است.

مدل QDA

برای انتخاب بهترین مدل qda ابتدا مدل‌ها را بر روی مجموعه داده برآورد برازش دادیم و سپس نرخ خطای طبقه‌بندی نادرست مدل‌ها را بر روی داده اعتبارسنجی محسابه کردیم.

جدول13- مدل qda

مدل	تفسیر مدل	نرخ خطای طبقه‌بندی نادرست داده اعتبارسنجی	نرخ خطای طبقه‌بندی نادرست داده برآورد
qda_1	در این مدل متغیر دیابت را در برابر تمام متغیرها بر روی داده برآورد برازش دادیم	0.1502591	0.1324675

اگر احتمال‌ها را برابر با 0.5 در نظر بگیریم خطاهای مدل ما به صورتی که در جدول 14 آمده است می‌شود.

جدول14- مدل qda

مدل	تفسیر مدل	نرخ خطای طبقه‌بندی نادرست داده اعتبارسنجی	نرخ خطای طبقه‌بندی نادرست داده برآورد
Qda_2	در این مدل متغیر دیابت را در برابر تمام متغیرها بر روی داده برآورد برازش دادیم	0.1450777	0.1311688

مدل بیز ساده

ابتدا مدل را بر روی مجموعه داده برآورد برازش دادیم و سپس خطاهارا محاسبه کردیم که در جدول 10 آمده است.

جدول15- مدل بیزساده

مدل	تفسیر مدل	نرخ خطای طبقه‌بندی نادرست داده اعتبارسنجی	نرخ خطای طبقه‌بندی نادرست داده برآورد
navie_bayse	در این مدل متغیر دیابت را در برابر تمام متغیرها بر روی داده برآورد برازش دادیم	0.134715	0.125974

در انتها تمامی مدل‌ها را در کنار یکدیگر گذاشته و نتایج را بررسی و بهترین مدل را انتخاب می‌کنیم.

جدول15- مقایسه مدل‌ها

مدل	نرخ خطای طبقه‌بندی نادرست داده برآورد	نرخ خطای طبقه‌بندی نادرست داده اعتبارسنجی
KNN	0.1129870	0.0932642
Tree	0.0636363	0.1191709
Random forest	0.0880829	0.0246753
log_reg_1	0.1142857	0.1142857
Lda_1	0.1181818	0.1243523
Lda_2	0.1181818	0.1191709
Qda_1	0.1324675	0.1502590
Qda_2	0.1311688	0.1450777
naive_baise	0.1259740	0.1347150

با توجه به جدول بالا بهترین مدل ما random forest می‌باشد اما، ما فقط می‌خواستیم دقت این مدل را نشان دهیم و به کمک آن متغیرهای مهم را نمایش بدهیم و آن را در ارزیابی نهایی دخالت نمی‌دهیم.

4.2-معرفی روش پیشنهادی – k-fold cross validation

برای دقت بالاتر به کمک k-fold cross validation مدل‌ها را ارزیابی می‌کنیم و مجموعه آموزشی را به 10 قسمت تقریبا مساوی تقسیم کرده ایم. حال هر بار یک قسمت را به عنوان اعتبارسنجی و 9 قسمت دیگر را برآورد درنظر می‌گیریم و مدل‌ها را اجرا می‌کنیم. میانگین خطا طبقه‌بندی را در جدول 16 مشاهده می‌کنید(توضیحات بیشتر در فایل کد).

جدول16- میانگین خطای طبقه‌بندی مدل‌ها

مدل	نرخ خطای طبقه‌بندی نادرست داده اعتبارسنجی	مدل	نرخ خطای طبقه‌بندی نادرست داده اعتبارسنجی
Tree	0.1101267	KNN	0.1287586
log_reg_1	0.1235288	Qda_1	0.1392075
Lda_1	0.1277169	Qda_2	0.1381551
Lda_2	0.1246134	naive_baise	0.1298110

از آنجایی که کلاس‌های پیش‌بینی تفاوت بسیار کمی داشتند برای همین مدل های lda_1 و lda_2 به یکدیگر و مدل های qda_1 و qda_2 به یکدیگر در خطای طبقه‌بندی نزدیک هستند.

با توجه به جدول 16 بهترین مدل درخت تصمیم است. حال آن را بر روی مجموعه آموزشی برازش می‌دهیم و خطای آن را بر روی مجموعه آزمون می‌سنجیم.

جدول17- نتیجه مدل نهایی

مدل	تفسیر مدل	نرخ خطای طبقه‌بندی نادرست داده در داده آموزشی	نرخ خطای طبقه‌بندی نادرست داده در داده آزمون
Final_model	در این مدل متغیر دیابت را در برابر تمام متغیرها بر روی داده آموزشی برازش دادیم	0.060228	0.153527

خطای مدل در مجموعه آزمون 0.15 می‌باشد که نشان‌دهنده 15 درصد خطا در پیش‌بینی‌های انجام شده است. مقادیر واقعی در برابر پیش‌بینی شده آن به صورت زیر است.

جدول18- مقادیر واقعی در برابر مقادیر پیش‌بینی شده توسط مدل نهایی

	مقادیر واقعی		
		0	1
مقادیر پیش‌بینی شده	0	100	17
	1	20	104

با توجه به مقادیر پیش‌بینی، 17 درصد خطا در پیش‌بینی وضعیت افراد سالم مشاهده می‌شود اما این مقدار در افراد دیابتی 14 درصد می‌باشد.

5-نتیجه‌گیری و جمع‌بندی

یادگیری ماشین در مباحث پزشکی دارای کاربردهای فراوانی است و در زمینه‌های مختلفی از تشخیص بیماری تا پیش‌بینی نتایج درمانی به کار گرفته می‌شود. در این گزارش به بررسی و پیش‌بینی بیماری دیابت در افراد پرداختیم. با استفاده از مدل‌های یادگیری ماشین، می‌توان با هزینه کمتر و سرعت بالا به کشف بیماری و وضعیت مریضان پرداخت. اما نکته قابل توجه در این زمینه، این است که قیمت خطا مدل مبلغ فروش خانه و ماشین و... نمی‌باشد و در این مسایل قیمت خطای آن، جان یک انسان است پس باید مدلی ساخت که قابل اتکا باشد و خطای بسیار کمی داشته باشد.

6-منابع

[1] Diabetes prediction dataset

https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset