پیش‌بینی دیابت با استفاده از مدل‌های یادگیری ماشین بانظارت

خلاصه

بیماری دیابت یکی از شایع‌ترین، خطرناک‌ترین و پرهزینه‌ترین بیماری‌های حال حاضر دنیا است که با نرخ هشداردهنده‌ای در حال افزایش است. استفاده از روش‌های داده‌کاوی می‌تواند به تشخیص زودهنگام دیابت کمک کند که باعث جلوگیری از پیشرفت این بیماری و بسیاری از عوارض آن مانند بیماری‌های قلبی و عروقی، مشکلات بینایی و نارسایی‌های کلیوی می‌شود .در این تحقیق از کتابخانه scikit-learn برای مدل‌سازی، به‌منظور دسته‌بندی بیماران به دو گروه دیابتی و غیر دیابتی استفاده شده است. داده‌های موردنیاز این تحقیق از پایگاه داده سایت kaggle استخراج شده است که شامل رکوردهای ۱۰۰،۰۰۰ مراجعه‌کننده می‌باشد. این داده‌ها شامل ویژگی‌های جنسیت، سن، فشارخون، بیماری قلبی، سابقه سیگار کشیدن، شاخص توده بدنی (BMI)، سطح قند خون سه ماهه (HbA1c) و سطح قند خون ناشتا هستند .پس از پیش‌پردازش و مدل‌سازی با استفاده از تکنیک‌های مختلف طبقه‌بندی، مدل Support Vector Machine (SVM) با دستیابی به ۹۲٪ فراخوانی (Recall) در تشخیص بیماری و دقت کل ۸۶٪، به‌عنوان مناسب‌ترین مدل برای اهداف غربالگری تشخیص داده شد.

۱. مقدمه

دیابت یک بیماری مزمن است که روی چگونگی تولید و استفاده انسولین در بدن اثر می‌گذارد. اگر سلول‌های بدن در مقابل انسولین مقاومت نشان دهند و یا اگر بدن به‌اندازه کافی انسولین نسازد، عملکرد بدن دچار اختلال خواهد شد. اشخاصی که به بیماری دیابت مبتلا هستند معمولاً تا زمانی که قند خونشان به بیش از دو برابر میزان نرمال نرسد، آثار فیزیکی چشمگیری مشاهده نمی‌کنند. در این حالت حتی برای کسانی که دیابت ندارند علائم بیماری کم‌وبیش مشابه است. مسئلهٔ قابل‌توجه در حوزهٔ پزشکی آن است که هنگامی‌که تعداد پارامترهای موردبررسی زیاد باشد، تشخیص صحیح حتی برای یک متخصص هم دشوار می‌شود. از طرفی کشف الگوهای پنهان و ارتباط میان پارامترهای تأثیرگذار بر بیماری به روش‌های سنتی تقریباً ناممکن است؛ بنابراین هدف اصلی این تحقیق کشف ارتباط میان عوامل تأثیرگذار و ایجاد مدلی هوشمند برای پیش‌بینی بیماری است.

در این پژوهش ۴ نوع دیابت مدنظر است: دیابت نوع یک (دیابت جوانی) که ۱۰ تا ۱۵ درصد موارد را شامل می‌شود؛ دیابت نوع دو (دیابت بزرگ‌سالان) که ۸۵ تا ۹۰ درصد موارد را شامل شده و بیشتر در افراد چاق بالای ۳۰ سال دیده می‌شود؛ دیابت حاملگی که گذراست و دیابت ناشی از علل متفرقه (جراحی، داروها و...). در این پژوهش از الگوریتم‌های طبقه‌بندی Gaussian Naive Bayes، Kneighbors، Support Vector Machine، Logistic Regression، AdaBoost، Decision Tree و Multi-layer Perceptron برای ایجاد مدل و شناسایی افراد مبتلا استفاده شده است.

۲. مفاهیم پایه

۱.۲ نقشه‌های حرارتی همبستگی

نقشه حرارتی همبستگی یک ابزار مفید برای نمایش گرافیکی چگونگی ارتباط دو ویژگی با یکدیگر است. بسته به نوع داده ویژگی‌ها، باید از روش‌های محاسبه ضریب همبستگی مناسب استفاده کنیم. به‌عنوان‌مثال می‌توان به ضریب همبستگی پیرسون اشاره کرد. ضریب همبستگی پیرسون معیاری از همبستگی خطی بین دو مجموعه داده است. نسبت بین کوواریانس دو متغیر و حاصل‌ضرب انحراف معیار آن‌هاست؛ بنابراین اساساً یک اندازه‌گیری نرمال شده از کوواریانس است، به‌طوری‌که نتیجه همیشه مقداری بین ۱− و ۱+ دارد.

۲.۲ معیارهای ارزشیابی

یک ماتریس درهم‌ریختگی چیدمان جدولی خاصی است که اجازه تجسم عملکرد یک الگوریتم یادگیری نظارت‌شده را می‌دهد. هر ردیف از ماتریس نشان‌دهنده نمونه‌ها در یک کلاس واقعی است درحالی‌که هر ستون نشان‌دهنده نمونه‌ها در یک کلاس پیش‌بینی‌شده است. جدول زیر مثالی از یک ماتریس درهم‌ریختگی برای یک طبقه‌بندی دودویی است که اصطلاحات/معیارهای دیگر را می‌توان از آن استخراج کرد.

	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

توضیحات	معنی	عبارت
موارد مثبتی که به‌عنوان مثبت پیش‌بینی شده‌اند.	مثبت صحیح	TP
موارد منفی که به‌عنوان مثبت پیش‌بینی شده‌اند.	مثبت کاذب	FP
موارد منفی که به‌عنوان منفی پیش‌بینی شده‌اند.	منفی صحیح	TN
موارد مثبتی که به‌عنوان منفی پیش‌بینی شده‌اند.	منفی کاذب	FN

۱. Accuracy: اندازه‌گیری تعداد مواردی که به‌درستی توسط مدل شناسایی/پیش‌بینی شده‌اند، یعنی پیش‌بینی صحیح تقسیم‌بر حجم کل نمونه.

$$\frac{TP+TN}{TP+TN+FP+FN}$$

۲. Recall: نرخ موارد مثبت صحیح را اندازه‌گیری می‌کند، یعنی چند مورد از موارد مثبت واقعی توسط مدل به‌عنوان مثبت شناسایی/پیش‌بینی می‌شوند.

$$\frac{TP}{(TP+FN)}$$

۳. Precision: اندازه‌گیری می‌کند که چه تعداد از موارد مثبت پیش‌بینی‌شده درواقع مثبت هستند.

$$\frac{TP}{(TP+FP)}$$

٤. F1-Score: Precision و Recall مدل را ترکیب می‌کند و به‌عنوان میانگین هارمونیک Precision و Recall مدل‌ها تعریف می‌شود.

$$2\times\frac{recall\times precision}{recall+precision}$$

۵. منحنی‌های ROC: منحنی مشخصه عملیاتی گیرنده (ROC)، یک نمودار گرافیکی است که عملکرد الگوریتم طبقه‌بندی دودویی را به‌عنوان تابعی از نرخ مثبت صحیح و نرخ مثبت کاذب نشان می‌دهد.

۳. روش انجام تحقیق

۱.۳ کاوش ویژگی‌ها و هدف

۱.۱.۳ خلاصه آماری

پس از بارگذاری دیتاست اولیه (df0) و نمایش ۱۰ سطر اول (جدول ۱)، ستون‌های موجود و نوع داده‌های آن‌ها با دستورات df0.columns و df0.dtypes بررسی شدند.

جدول ۱ – نمایش ۱۰ سطر اول دیتافریم df0 (داده‌های خام).

	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
0	Female	80.0	0	1	never	25.19	6.6	140	0
1	Female	54.0	0	0	No Info	27.32	6.6	80	0
2	Male	28.0	0	0	never	27.32	5.7	158	0
3	Female	36.0	0	0	current	23.45	5.0	155	0
4	Male	76.0	1	1	current	20.14	4.8	155	0
5	Female	20.0	0	0	never	27.32	6.6	85	0
6	Female	44.0	0	0	never	19.31	6.5	200	1
7	Female	79.0	0	0	No Info	23.86	5.7	85	0
8	Male	42.0	0	0	never	33.64	4.8	145	0
9	Female	32.0	0	0	never	27.32	5.0	100	0

با اجرای دستور df0["diabetes"].value_counts()، توزیع اولیه برچسب هدف نشان داد که ۹۱,۵۰۰ نمونه سالم (کلاس ۰) و ۸,۵۰۰ نمونه دیابتی (کلاس ۱) وجود دارد؛ این عدم‌توازن شدید کلاس‌ها (نسبت تقریبی ۱۰:۱) یکی از چالش‌های اصلی مدل‌سازی است که در بخش ۲.۲.۳ با استفاده از تکنیک SMOTE برطرف شده است.

برای ویژگی‌های عددی (age, bmi, HbA1c_level, blood_glucose_level)، دستور

df1[["age", "bmi", "HbA1c_level", "blood_glucose_level"]].describe()

اعمال شد و خلاصه آماری حاصل در جدول ۲ درج گردید.

جدول ۲ — خلاصه آماری ویژگی‌های عددی دیتافریم df1.

	age	bmi	HbA1c_level	blood_glucose_level
count	96128.000000	96128.000000	96128.000000	96128.000000
mean	41.796617	27.321450	5.532633	138.218001
std	22.463329	6.767811	1.073225	40.911190
min	0.080000	10.010000	3.500000	80.000000
25%	24.000000	23.400000	4.800000	100.000000
50%	43.000000	27.320000	5.800000	140.000000
75%	59.000000	29.860000	6.200000	159.000000
max	80.000000	95.690000	9.000000	300.000000

این جدول نشان می‌دهد که میانگین سن افراد ۴۱.۸ سال، میانگین BMI حدود ۲۷.۳، میانگین HbA1c حدود ۵.۵٪ و میانگین قند خون ناشتا ۱۳۸ mg/dL است.

۲.۱.۳ توزیع ویژگی‌ها و هدف

در شکل ۱، هیستوگرام تمامی ویژگی‌ها (پس از کدگذاری) به تصویر کشیده شده است. نکات کلیدی استخراج‌شده از این نمودارها عبارتند از:

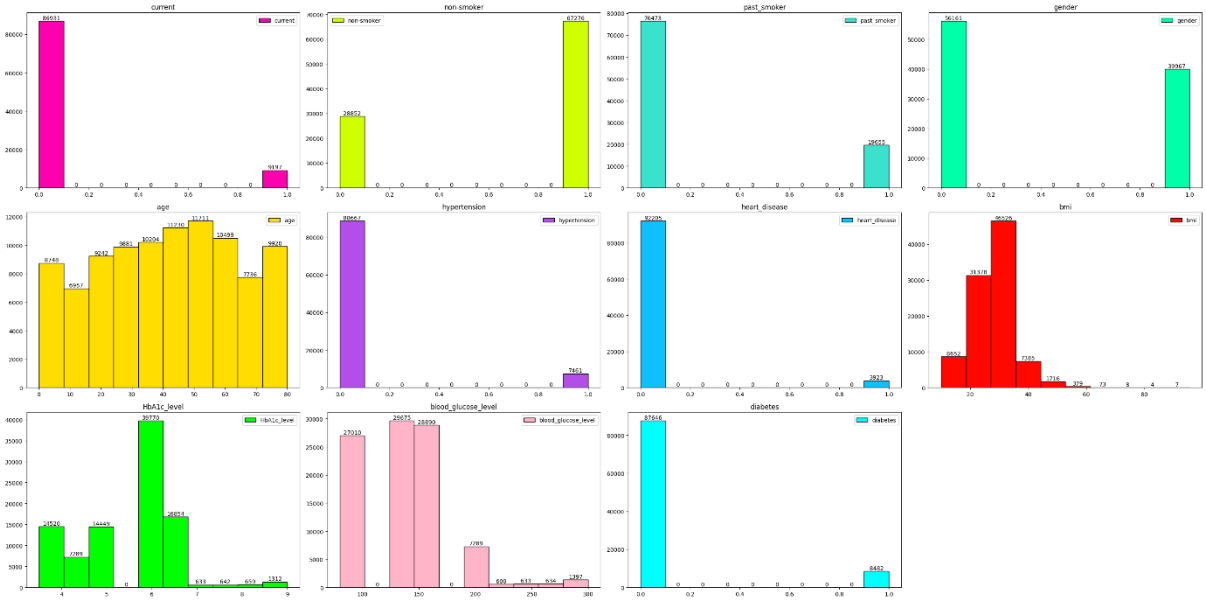
مصرف سیگار: ۸۶,۹۳۱ نفر در حال حاضر سیگار نمی‌کشند (non-smoker)، ۹,۱۹۷ نفر سیگاری فعال (current)، و ۲۸,۸۵۲ نفر سابقه مصرف دارند (past_smoker).

جنسیت: ۵۶,۱۶۱ نفر زن و ۳۹,۹۶۷ نفر مرد.

فشارخون: ۸۸,۶۶۷ نفر فشارخون ندارند؛ ۷,۴۶۱ نفر دارند.

بیماری قلبی: ۹۲,۲۰۵ نفر سابقه بیماری قلبی ندارند؛ ۳,۹۲۳ نفر دارند.

توزیع هدف: ۸۷,۶۴۶ نفر سالم و ۸,۴۸۲ نفر دیابتی (پس از حذف سطرهای تکراری و null).



شکل ۱ — هیستوگرام توزیع تمامی ویژگی‌ها و هدف در دیتافریم df1.

۲.۳ پیش‌پردازش داده‌ها

۱.۲.۳ پاک‌سازی و کدگذاری

حذف تکراری و null: سطرهای تکراری (()duplicated) و سطرهای حاوی مقادیر گمشده (()isnull) حذف شدند.

کدگذاری جنسیت: ۱۸ سطر با جنسیت "Other" (به‌دلیل تعداد ناچیز) حذف شدند و سپس Female=0 و Male=1 تعیین شد (جدول ۳).

جدول ۳ – نمایش ۱۰ سطر اول دیتافریم پس از کدگذاری ویژگی جنسیت.

	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
0	0	80.0	0	1	non-smoker	25.19	6.6	140	0
1	0	54.0	0	0	non-smoker	27.32	6.6	80	0
2	1	28.0	0	0	non-smoker	27.32	5.7	158	0
3	0	36.0	0	0	current	23.45	5.0	155	0
4	1	76.0	1	1	current	20.14	4.8	155	0
5	0	20.0	0	0	non-smoker	27.32	6.6	85	0
6	0	44.0	0	0	non-smoker	19.31	6.5	200	1
7	0	79.0	0	0	non-smoker	23.86	5.7	85	0
8	1	42.0	0	0	non-smoker	33.64	4.8	145	0
9	0	32.0	0	0	non-smoker	27.32	5.0	100	0

بازطبقه‌بندی مصرف سیگار: ۶ رسته اولیه (never, No Info, current, former, ever, not current) به ۳ رسته کلی تبدیل شدند:

non-smoker ← never + No Info
current ← current
past_smoker ← former + ever + not current

OneHotEncoding: رسته‌های جدید مصرف سیگار با OneHotEncoder به ۳ ستون باینری (current, non-smoker, past_smoker) تبدیل شدند و دیتافریم نهایی df1 تولید شد (جدول ۴).

جدول ۴ – نمایش ۱۰ سطر اول دیتافریم df1 پس از OneHotEncoding ویژگی سابقه مصرف سیگار.

	current	non-smoker	past_smoker	gender	age	hypertension	heart_disease	bmi	HbA1c_level	blood_glucose_level	diabetes
0	0.0	1.0	0.0	0.0	80.0	0.0	1.0	25.19	6.6	140.0	0.0
1	0.0	1.0	0.0	0.0	54.0	0.0	0.0	27.32	6.6	80.0	0.0
2	0.0	1.0	0.0	1.0	28.0	0.0	0.0	27.32	5.7	158.0	0.0
3	1.0	0.0	0.0	0.0	36.0	0.0	0.0	23.45	5.0	155.0	0.0
4	1.0	0.0	0.0	1.0	76.0	1.0	1.0	20.14	4.8	155.0	0.0
5	0.0	1.0	0.0	0.0	20.0	0.0	0.0	27.32	6.6	85.0	0.0
6	0.0	1.0	0.0	0.0	44.0	0.0	0.0	19.31	6.5	200.0	1.0
7	0.0	1.0	0.0	0.0	79.0	0.0	0.0	23.86	5.7	85.0	0.0
8	0.0	1.0	0.0	1.0	42.0	0.0	0.0	33.64	4.8	145.0	0.0
9	0.0	1.0	0.0	0.0	32.0	0.0	0.0	27.32	5.0	100.0	0.0

۲.۲.۳ تقسیم داده و متوازن‌سازی

Train/Test split: ۷۵٪ آموزش (۷۲,۰۹۶ نمونه) و ۲۵٪ آزمون (۲۴,۰۳۲ نمونه) با استفاده از train_test_split و stratify=Y.

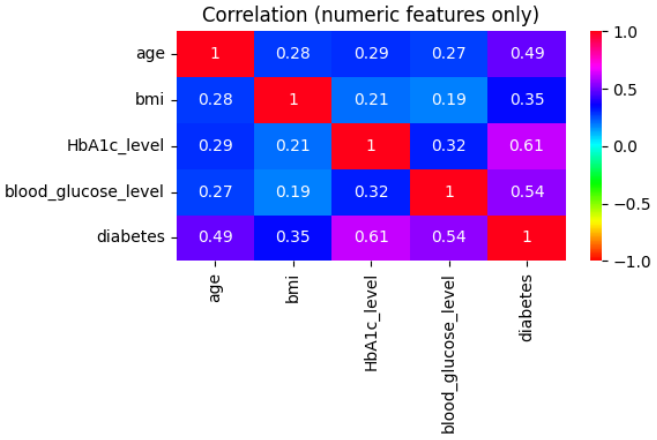
SMOTE: به‌دلیل نسبت تقریبی ۱۰:۱ کلاس‌ها، تکنیک SMOTE روی داده‌های آموزش اعمال شد و تعداد نمونه‌های آموزش از ۷۲,۰۹۶ به ۱۳۱,۴۶۸ افزایش یافت.

استانداردسازی: ویژگی‌های عددی (age, bmi, HbA1c_level, blood_glucose_level) با StandardScaler نرمال شدند.

۳.۳ تحلیل همبستگی و انتخاب ویژگی

۱.۳.۳ ماتریس همبستگی

در شکل ۲، ماتریس همبستگی پیرسون برای ویژگی‌های عددی و هدف رسم شده است.



شکل ۲ – ماتریس همبستگی (Correlation Heatmap) ویژگی‌های عددی با هدف diabetes.

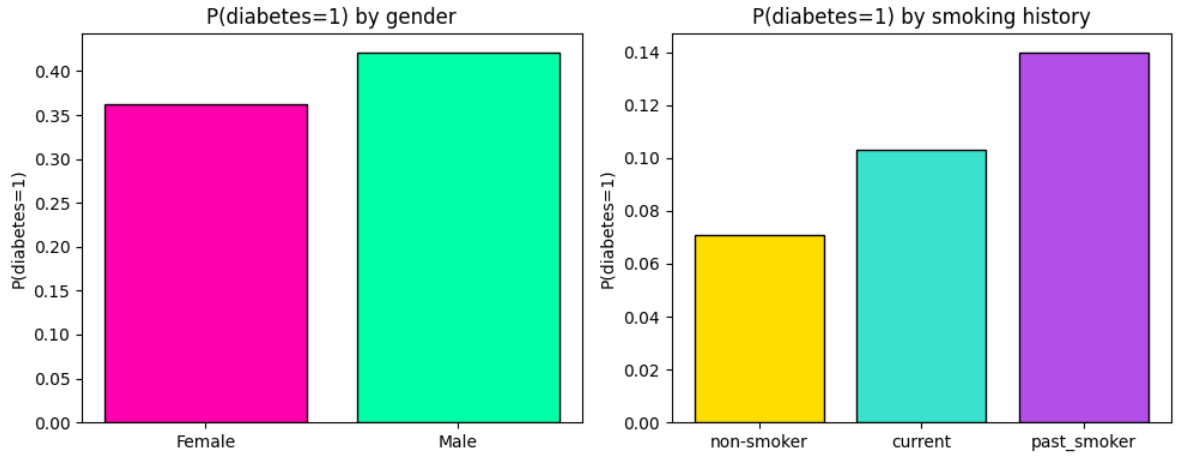
نتایج نشان می‌دهند:

HbA1c_level با ضریب همبستگی ۰.۶۱۱ بیشترین ارتباط خطی با دیابت را دارد.

blood_glucose_level با ضریب ۰.۵۳۹ در رتبه دوم قرار دارد.

Age (۰.۴۹۱) و bmi (۰.۳۴۸) نیز تأثیرگذار هستند اما ضعیف‌تر.

برای بررسی ارتباط ویژگی‌های رسته‌ای (جنسیت و سابقه مصرف سیگار) با هدف، احتمال ابتلا به دیابت (P(diabetes=1)) برای هر دسته محاسبه و در شکل ۳ نمایش داده شده است.



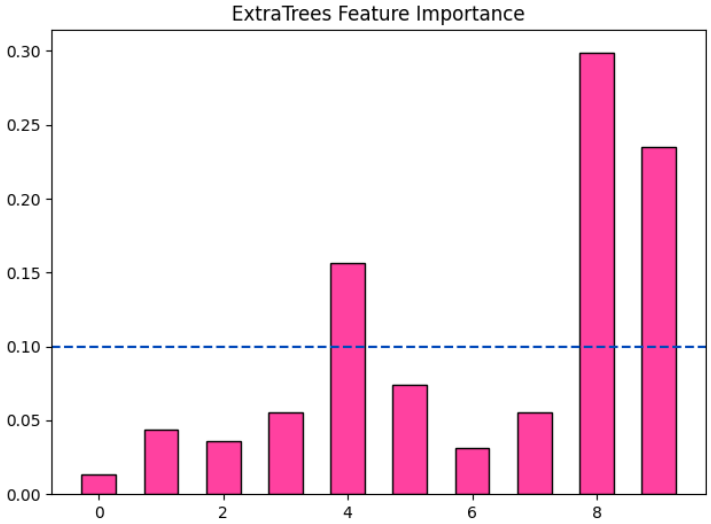
شکل ۳ – مقایسه احتمال ابتلا به دیابت بر اساس جنسیت (چپ) و سابقه مصرف سیگار (راست).

مشاهده می‌شود که:

- جنسیت: مردان (Male) احتمال ابتلای بالاتری (حدود ۰.۴۱) نسبت به زنان (Female، حدود ۰.۳۶) دارند.
- سابقه مصرف سیگار: افراد با سابقه مصرف سیگار (past_smoker) بیشترین احتمال ابتلا (حدود ۰.۱۴) و افراد غیرسیگاری (non-smoker) کمترین احتمال (حدود ۰.۰۷) را دارند.

۲.۳.۳ انتخاب ویژگی با ExtraTrees

برای کاهش ابعاد و حذف ویژگی‌های کم‌اهمیت، مدل ExtraTreesClassifier (با ۱۰۰ درخت) آموزش داده شد و امتیاز اهمیت (Feature Importance) هر ویژگی محاسبه گردید (شکل ۴).



شکل ۴ – نمودار اهمیت ویژگی‌ها (Feature Importance) با استفاده از ExtraTreesClassifier. خط افقی آبی نشان‌دهنده میانگین امتیازها است.

ویژگی‌هایی که امتیاز بالاتر از میانگین (۰.۱۰) داشتند، انتخاب شدند:

ویژگی ۴ (age): ۰.۱۵۶

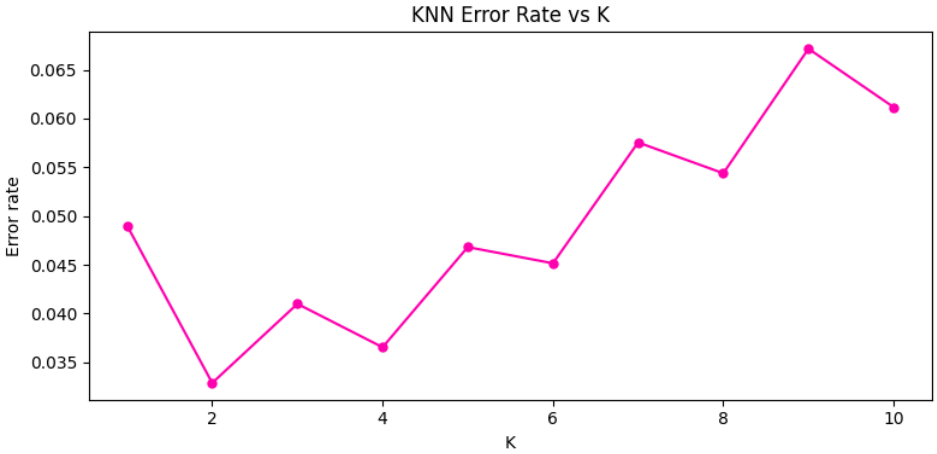
ویژگی ۸ (HbA1c_level): ۰.۲۹۹

ویژگی ۹ (blood_glucose_level): ۰.۲۳۵

در نتیجه، سه ویژگی age، HbA1c_level و blood_glucose_level به‌عنوان مهم‌ترین و مؤثرترین ویژگی‌ها بر هدف انتخاب شدند و فضای ویژگی از ۱۰ به ۳ بُعد کاهش یافت؛ Xte: (24032, 3) و Xtr: (131468, 3). نگاهی گذرا به شکل ۲ (ماتریس همبستگی) به‌طور واضح مشخص می‌کند که چرا این سه ویژگی نسبت به سایر ویژگی‌ها تأثیر بیشتری در تعیین کلاس هر نمونه دارند؛ زیرا بالاترین ضرایب همبستگی خطی با هدف diabetes را دارا هستند.

۴. نتایج و بحث

در این تحقیق ۷ طبقه‌بند Gaussian Naive Bayes، Kneighbors، Support Vector، Logistic Regression، AdaBoost، Decision Tree و Multi-layer Perceptron مورد استفاده قرار گرفته‌اند. تعدادی از این طبقه‌بندها به‌صورت پیش‌فرض مورد استفاده قرار گرفته‌اند و در تعدادی دیگر، پارامترها یا هایپرپارامترهای آن‌ها تنظیم یا انتخاب شده‌اند. به‌عنوان‌مثال، تعداد همسایه‌های (K) مؤثر بر کارایی طبقه‌بند Kneighbors در یک فرایند تکراری و به‌واسطه ارزیابی میانگین نرخ خطای ناشی از عدم برابری برچسب داده‌های آزمون و مقدار پیش‌بینی‌شده متناظر برای آن برچسب به دست می‌آید (شکل (۵)). با بررسی نرخ خطا برای مقادیر مختلف همسایگی، تعداد ۲ همسایه (K=2) به عنوان مقدار بهینه انتخاب شد که کمترین نرخ خطا را در پی داشت.



شکل ۵ – نحوهٔ مشخص کردن K همسایه نزدیک مورداستفاده در الگوریتم Kneighbors با توجه به کمترین میانگین نرخ خطا.

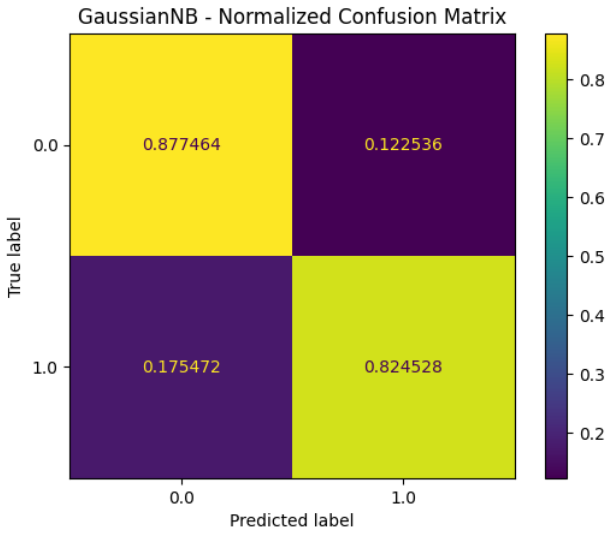
در ادامه، نتایج حاصل از آموزش و ارزیابی ۷ طبقه‌بند مختلف به ترتیب ارائه شده است. عملکرد هر یک از این مدل‌ها با استفاده از معیارهای Accuracy، Precision، Recall و ماتریس درهم‌ریختگی مورد بررسی قرار گرفته است تا در نهایت با مقایسه نتایج، کارآمدترین مدل جهت تشخیص بیماری انتخاب گردد.

```
Gaussian Naive Bayes Classifier
GaussianNB - Train acc: 0.8630313079989047  Test acc: 0.8727946071904128
              precision    recall  f1-score   support

         0.0       0.98      0.88      0.93     21912
         1.0       0.39      0.82      0.53      2120

    accuracy                           0.87     24032
   macro avg       0.69      0.85      0.73     24032
weighted avg       0.93      0.87      0.89     24032


=== Overall classifier results so far ===
        model  train_acc  test_acc  relative_to_best
0  GaussianNB   0.863031  0.872795               1.0
==========================================
```



شکل ۶ – ماتریس درهم‌ریختگی مدل مبتنی بر Gaussian Naive Bayes.

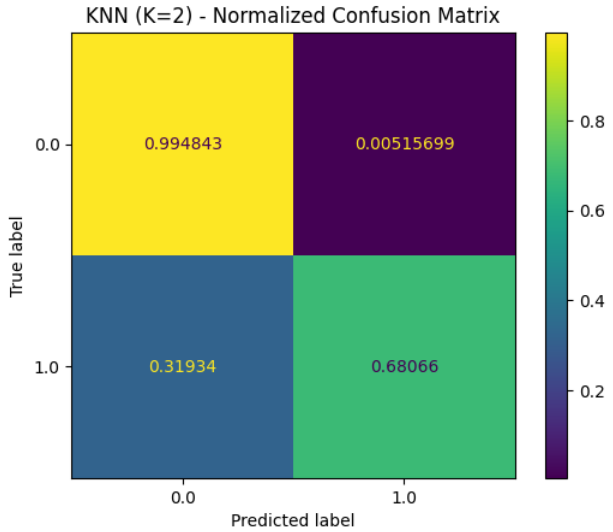
KNeighbors Classifier

```
Best K according to error rate: K=2
KNN (K=2) - Train acc: 0.9641  Test acc: 0.9671
              precision    recall  f1-score   support

         0.0       0.97      0.99      0.98     21912
         1.0       0.93      0.68      0.79      2120

    accuracy                           0.97     24032
   macro avg       0.95      0.84      0.88     24032
weighted avg       0.97      0.97      0.96     24032


=== Overall classifier results so far ===
        model  train_acc  test_acc  relative_to_best
0  GaussianNB   0.863031  0.872795             0.902
1    KNN(K=2)   0.964128  0.967127             1.000
========================================
```



شکل ۷ – ماتریس درهم‌ریختگی مدل مبتنی بر Kneighbors.

Support Vector Machine

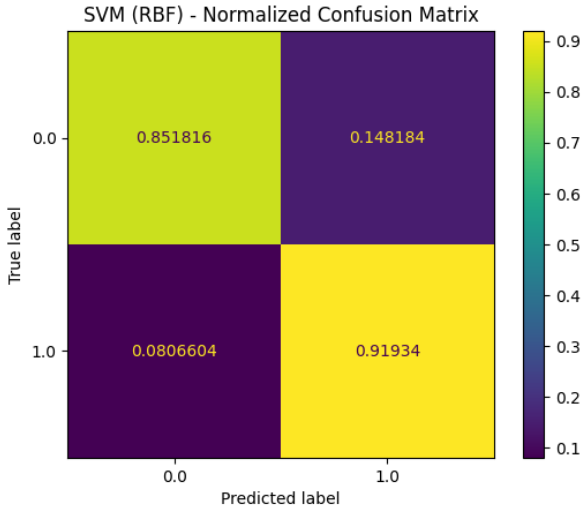
```
SVM (RBF) - Train acc: 0.8955106946177016  Test acc: 0.8577729693741678
```

```
              precision    recall  f1-score   support

         0.0       0.99      0.85      0.92     21912
         1.0       0.38      0.92      0.53      2120

    accuracy                           0.86     24032
   macro avg       0.68      0.89      0.72     24032
weighted avg       0.94      0.86      0.88     24032


=== Overall classifier results so far ===
        model  train_acc  test_acc  relative_to_best
0  GaussianNB   0.863031  0.872795             0.902
1    KNN(K=2)   0.964128  0.967127             1.000
2   SVM (RBF)   0.895511  0.857773             0.887
========================================
```



شکل ۸ – ماتریس درهم‌ریختگی مدل مبتنی بر Support Vector.

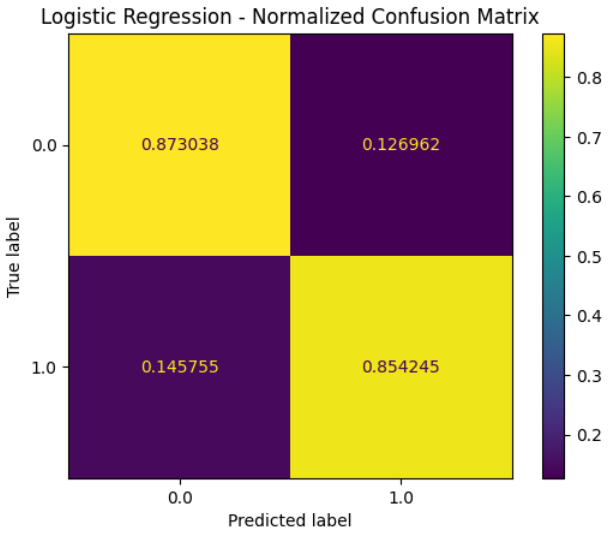
Logistic Regression

```
Logistic Regression - Train acc: 0.8767076398819484  Test acc: 0.8713798268974701
              precision    recall  f1-score   support

         0.0       0.98      0.87      0.93     21912
         1.0       0.39      0.85      0.54      2120

    accuracy                           0.87     24032
   macro avg       0.69      0.86      0.73     24032
weighted avg       0.93      0.87      0.89     24032


=== Overall classifier results so far ===
                 model  train_acc  test_acc  relative_to_best
0           GaussianNB   0.863031  0.872795             0.902
1             KNN(K=2)   0.964128  0.967127             1.000
2            SVM (RBF)   0.895511  0.857773             0.887
3  Logistic Regression   0.876708  0.871380             0.901
========================================
```



شکل ۹ – ماتریس درهم‌ریختگی مدل مبتنی بر Logistic Regression.

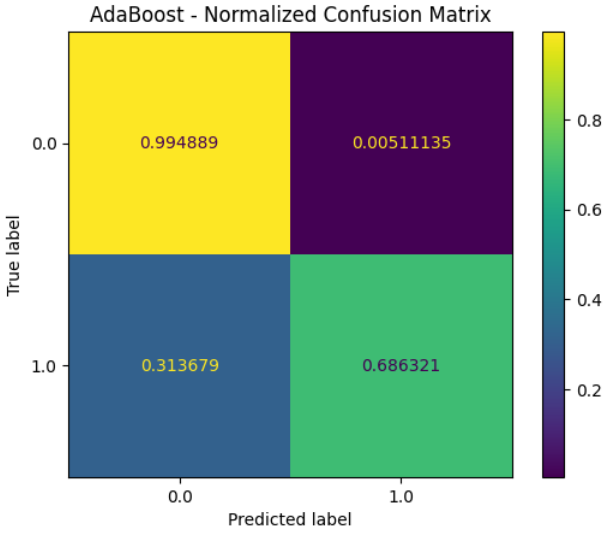
AdaBoost Classifier

```
AdaBoost - Train acc: 0.9664633218730033  Test acc: 0.9676681091877497
              precision    recall  f1-score   support

         0.0       0.97      0.99      0.98     21912
         1.0       0.93      0.69      0.79      2120

    accuracy                           0.97     24032
   macro avg       0.95      0.84      0.89     24032
weighted avg       0.97      0.97      0.97     24032


=== Overall classifier results so far ===
                model  train_acc  test_acc  relative_to_best
0          GaussianNB   0.863031  0.872795             0.902
1            KNN(K=2)   0.964128  0.967127             0.999
2           SVM (RBF)   0.895511  0.857773             0.886
3  Logistic Regression   0.876708  0.871380             0.900
4            AdaBoost   0.966463  0.967668             1.000
========================================
```



شکل ۱۰ – ماتریس درهم‌ریختگی مدل مبتنی بر AdaBoost.

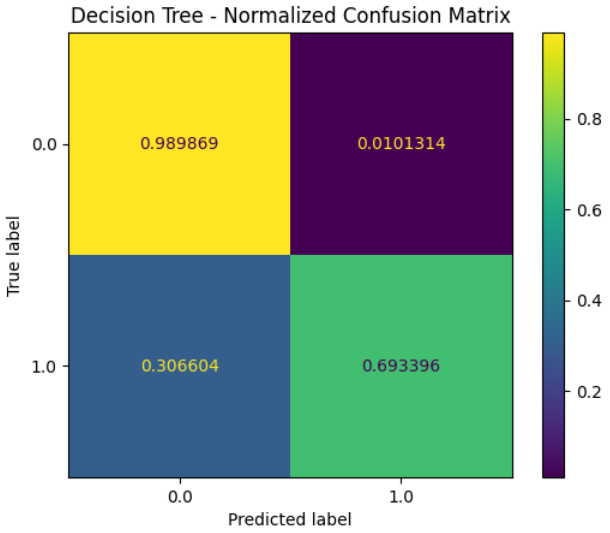
Decision Tree Classifier

```
Decision Tree - Train acc: 0.981904341740956  Test acc: 0.9637150466045273
              precision    recall  f1-score   support

         0.0       0.97      0.99      0.98     21912
         1.0       0.87      0.69      0.77      2120

    accuracy                           0.96     24032
   macro avg       0.92      0.84      0.88     24032
weighted avg       0.96      0.96      0.96     24032


=== Overall classifier results so far ===
                model  train_acc  test_acc  relative_to_best
0          GaussianNB   0.863031  0.872795             0.902
1            KNN(K=2)   0.964128  0.967127             0.999
2           SVM (RBF)   0.895511  0.857773             0.886
3  Logistic Regression   0.876708  0.871380             0.900
4            AdaBoost   0.966463  0.967668             1.000
5       Decision Tree   0.981904  0.963715             0.996
========================================
```



شکل ۱۱ – ماتریس درهم‌ریختگی مدل مبتنی بر Decision Tree.

Multi-layer Perceptron classifier

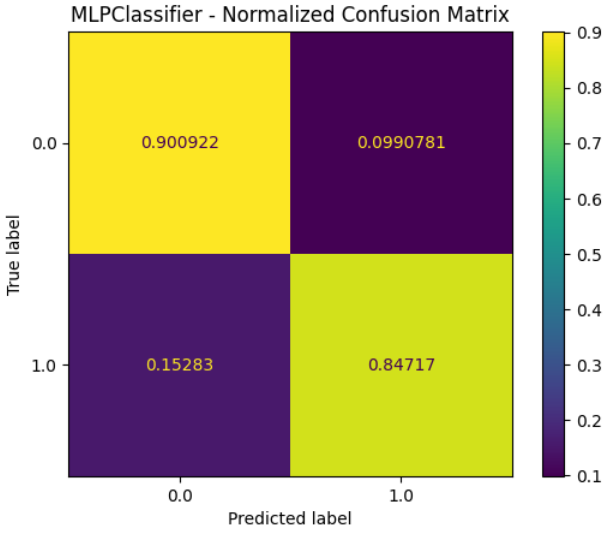
```
MLPClassifier - Train acc: 0.9112255453798643  Test acc: 0.8961800932090546
              precision    recall  f1-score   support

         0.0       0.98      0.90      0.94     21912
         1.0       0.45      0.85      0.59      2120

    accuracy                           0.90     24032
   macro avg       0.72      0.87      0.77     24032
weighted avg       0.94      0.90      0.91     24032


=== Overall classifier results so far ===
                model  train_acc  test_acc  relative_to_best
0          GaussianNB   0.863031  0.872795             0.902
1            KNN(K=2)   0.964128  0.967127             0.999
2           SVM (RBF)   0.895511  0.857773             0.886
3  Logistic Regression   0.876708  0.871380             0.900
```

```
4            AdaBoost   0.966463  0.967668            1.000
5       Decision Tree   0.981904  0.963715            0.996
6       MLPClassifier   0.911226  0.896180            0.926
========================================
```



شکل ۱۲ – ماتریس درهم‌ریختگی مدل مبتنی بر Multi-layer Perceptron.

```
=== FINAL CLASSIFIER SUMMARY ===

=== Overall classifier results so far ===
                 model  train_acc  test_acc  relative_to_best
0          GaussianNB   0.863031  0.872795             0.902
1            KNN(K=2)   0.964128  0.967127             0.999
2           SVM (RBF)   0.895511  0.857773             0.886
3  Logistic Regression   0.876708  0.871380             0.900
4            AdaBoost   0.966463  0.967668             1.000
5       Decision Tree   0.981904  0.963715             0.996
6       MLPClassifier   0.911226  0.896180             0.926
========================================
```

۱.۴ جمع‌بندی و مقایسه عملکرد مدل‌ها

با تحلیل جامع نتایج به‌دست‌آمده از ۷ طبقه‌بند مورد مطالعه، تفاوت آشکاری میان «دقت کلی» و «قدرت تشخیص بیماری» در مدل‌های مختلف مشاهده می‌شود. اگرچه مدل‌هایی نظیر KNN و AdaBoost به دقت کلی (Accuracy) بسیار بالایی (حدود ۹۷٪) دست یافته‌اند، اما بررسی ماتریس درهم‌ریختگی آن‌ها نشان می‌دهد که در معیار حیاتی فراخوانی (Recall) عملکرد ضعیفی داشته‌اند (حدود ۶۸٪ تا ۶۹٪). این بدان معناست که این مدل‌ها علیرغم عملکرد خوب روی داده‌های سالم، تعداد قابل‌توجهی از بیماران واقعی را به اشتباه سالم تشخیص داده‌اند (منفی کاذب بالا) که در کاربردهای پزشکی خطای خطرناکی محسوب می‌شود.

در سوی دیگر، مدل‌های Logistic Regression و MLP با نرخ فراخوانی حدود ۸۵٪ عملکرد قابل‌قبولی در شناسایی بیماران داشتند، اما مدل Support Vector Machine (SVM) با اختلاف معناداری نسبت به سایر روش‌ها، توانست به نرخ فراخوانی ۹۲٪ دست یابد.

گفتنی است که مدل SVM دارای دقت کلی ۸۶٪ و نرخ دقت (Precision) نسبتاً پایین ۳۸٪ است؛ این یعنی مدل مذکور در برخی موارد افراد سالم را نیز مشکوک به دیابت تشخیص می‌دهد (مثبت کاذب). با این حال، در طراحی سامانه‌های غربالگری و کمک‌تشخیصی، استراتژی صحیح بر مبنای «به حداقل رساندن منفی‌های کاذب» استوار است. به بیان دیگر، هزینه تشخیص اشتباه یک فرد سالم به عنوان بیمار (که با یک آزمایش تکمیلی ساده رد می‌شود) بسیار کمتر از هزینه تشخیص ندادن یک فرد بیمار و پیشرفت بیماری در اوست.

بنابراین، مدل SVM به دلیل حساسیت بالا و توانایی شناسایی ۹۲ درصد از کل موارد دیابتی، به‌عنوان مطمئن‌ترین و کارآمدترین مدل جهت اهداف این پژوهش انتخاب گردید.

۵. نتیجه‌گیری و کارهای آتی

در این تحقیق، داده‌های ۱۰۰,۰۰۰ مراجعه‌کننده پس از پاک‌سازی، کدگذاری و متوازن‌سازی با تکنیک SMOTE، برای پیش‌بینی دیابت مورد استفاده قرار گرفتند. پس از اعمال کاهش ویژگی با روش ExtraTrees، سه ویژگی سن، سطح HbA1c و قند خون ناشتا به عنوان مؤثرترین فاکتورها شناسایی شدند. ارزیابی ۷ الگوریتم یادگیری ماشین نشان داد که مدل SVM با کرنل RBF، علی‌رغم داشتن دقت کلی (Accuracy) ۸۶٪ که پایین‌تر از مدل‌هایی مانند KNN (۹۷٪) است، توانست به فراخوانی (Recall) ۹۲٪ دست یابد. این بدان معناست که این مدل توانایی شناسایی ۹۲ درصد از کل بیماران دیابتی را دارد که در کاربردهای پزشکی و غربالگری، اولویت اصلی محسوب می‌شود.

یکی از مواردی که می‌تواند به این تحقیق و تحقیقات آتی درزمینهٔ کار روی داده‌های آزمایشگاهی کمک کند، ایجاد یک پرونده الکترونیکی جامع از وضعیت و سوابق پزشکی هر بیمار شامل پارامترهای فیزیکی مانند فشارخون، وزن، قد، میزان دور شکم و توده چربی در آزمایشگاه‌ها می‌باشد. با در اختیار داشتن این اطلاعات می‌توان به دقت مدل بالاتری به‌منظور پیش‌بینی ریسک بیماری و تشخیص دیابت دست یافت.

این مدل می‌تواند در مراکز درمانی و آزمایشگاه‌ها به‌عنوان یک سامانه کمک‌تشخیصی هوشمند مورد استفاده قرار گیرد تا با تحلیل پارامترهای موجود (مانند سن و سوابق)، ریسک ابتلا را در افرادی که علائم ظاهری ندارند، با دقت بالایی شناسایی کرده و آن‌ها را برای آزمایش‌های تکمیلی ارجاع دهد.

مراجع و منابع مورد استفاده

[1] A. Mujumdar and V. Vaidehi, "Diabetes prediction using machine learning algorithms," Procedia Computer Science, vol. 165, pp. 292-299, 2019.

[2] T. Sasada, Z. Liu, T. Baba, K. Hatano, and Y. Kimura, "A resampling method for imbalanced datasets considering noise and overlap," Procedia Computer Science, vol. 176, pp. 420-429, 2020.

[3] https://www.kaggle.com/code/gastonpascaltonguino/diabetes-dataset-eda-modeling