به نام خدا

دانشگاه تربیت دبیر شهید رجایی

دانشکده کامپیوتر

پروژه پایانی درس داده کاوی

نام دانشجو:

معصومه ایرانپور

شماره دانشجویی:

4001231008

استاد درس

دکتر دانشپور

لینک مخزن گیت هاب پروژه: masumeirp/data-mining-

فهرست مطالب

مقدمه

امروزه استفاده از روش‌های مبتنی بر هوش مصنوعی در پزشکی یکی از مهم‌ترین و داغ‌ترین مباحث برای محققین است. استفاده از این روش‌ها می‌تواند در تشخیص، درمان و یا پیشگیری بیماری‌ها کمک به‌سزایی کند و احتمال بروز خطای فردی را در این مراحل کاهش دهد.

آمارهای فدراسیون جهانی دیابت نشان می‌دهد که دیابت به معضل بهداشتی- درمانی و اجتماعی در کل جهان تبدیل شده است زیرا شیوع آن طی ۲۵ سال گذشته حدود ۴ برابر افزایش داشته و پیش بینی ۲۵ سال آینده، از افزایش ۳ برابری حکایت دارد. پیش‌بینی دیابت به عوامل مختلفی از جمله سن، شاخص توده‌ی بدنی، ژنتیک و... بستگی دارد اما استخراج ویژگی و طبقه‌بندی این داده‌ها با استفاده از الگوریتم‌های کلاسیک کاری سخت و یا غیرممکن خواهد بود. در این پروژه با ابتدا داده‌ها پیش‌پردازش شده و عملیات طبقه‌بندی و خوشه‌بندی روی آن‌ها انجام خواهد شد.

این پروژه بر بستر گوگل کلب و به زبان پایتون نوشته شده است.

بخش اول-پیش‌پردازش

نتیجه‌ی پروژه‌های پردازشی وابسته به دو بخش است: پیش‌پردازش و الگوریتم. تغییر در هرکدام از این بخش‌ها می‌تواند ما را به نتیجه‌ی مناسب نزدیک کرده و یا کاملا به سمت اشتباه پیش ببرد.

ابتدا به مراحل مختلف پاکسازی داده‌ها[1] می‌پردازیم. هنگام جمع‌آوری داده‌ها مشکلات مختلفی ممکن است بوجود بیاید. این مشکلات شامل مواردی مانند خطای انسانی، مشکالت سیستمی یا نقص در فرایند جمع‌آوری می‌باشند.

1. مقدار گمشده[2]

روش‌های مختلفی برای پر کردن مقادیر گمشده وجود دارد که از جمله‌های آن‌ها می‌توان به جایگزین کردن با مقدار میانگین و میانه، استفاده از مدل‌های پیش‌بینی و یا حذف این ردیف داده‌ها استفاده کرد.

با فراخوانی داده‌ها در پایتون و مشاهده می‌شود که داده‌ی موجود دارای 100001 ردیف می‌باشد. پس از آن تعداد مقادیر خالی ستون‌ها را چاپ می‌کنیم.

Count of Missing Values

	0
gender	0
age	2
hypertension	0
heart_disease	0
smoking_history	1
bmi	0
HbA1c_level	1
blood_glucose_level	0
diabetes	0

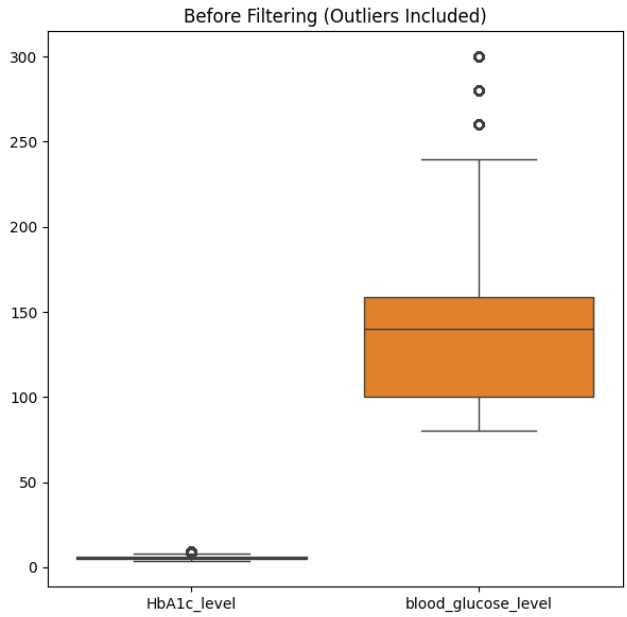
شکل 1 مقدار گمشده

همانطور که مشاهده می‌شود در مجموع 4 داده‌ی گمشده وجود دارد. به علت اینکه غیر از سابقه‌ی استعمال سیگار 3 مورد باقیمانده مهم‌ترین ویژگی‌های مورد نظر هستند و تعداد کل آن‌ها در مقایسه با تعداد داده‌ها عدد بسیار کوچکی است، برای جلوگیری از خطای احتمالی ستون‌های دارای مقادیر خالی را از داده حذف می‌کنیم.

2. مقادیر خارج از محدوده[3]

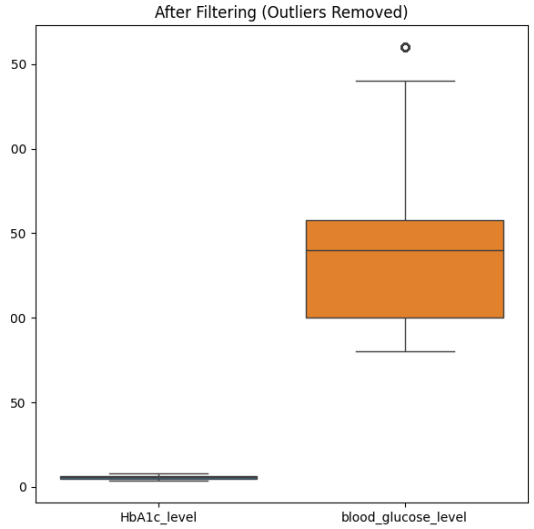
مقادیر خارج از محدوده نیز به دلایل مختلفی ایجاد می‌شوند. مقادیر خارج از محدوده میتوانند نتایج به دست آمده را دستخوش تغییراتی کنند که باعث گمراهی ما شود. برای مثال اگر میانگین گیری بین داده ها انجام دهیم و رنج همه ی اعداد یکی باشد به جز یک عدد، آن یک عدد نتیجه را منحرف خواهد کرد. معمولا برای رسیدگی به مقادیر خارج از محدوده، داده‌هایی که ازانحراف معیار داده‌ها دور هستند را حذف می‌کنند. برای پیدا کردن این داده‌ها z-score را محاسبه کرده و آستانه را برابر با ± 3 قرار می‌دهیم.

[1] Data Cleaning
[2] Missing Data
[3] Outliers



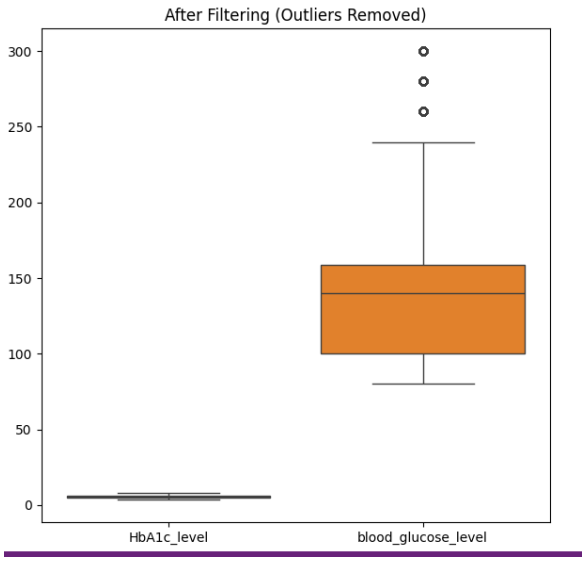
شکل 2 نمودار جعبه‌ای داده‌ها قبل از فیلتر شدن

همانطور که مشاهده می‌شود مقدار قند خون می‌تواند در مقادیر گسترده‌ای باشد. بنابراین با حذف مقادیر داده‌های پرت در آن با استفاده از z-score، تنها افرادی که احتمال دیابت در آن‌ها بالاتر است حذف می‌شوند و نمودار جعبه‌ای به شکل زیر در می‌آید.



شکل 3 نمودار جعبه‌ای داده‌ها بعد از فیلتر شدن با z-score

به همین دلیل فقط مقادیر پرت HbA1c را با این روش حذف می‌کنیم . مقادیر قند خون را به بین 40 تا 500 فیلتر می‌کنیم. نمودار جعبه‌ای داده‌ی فیلتر شده به شکل زیر تبدیل می‌شود.



شکل 4 نمودار جعبه‌ای نهایی

پس از فیلتر کردن تعداد داده‌های موجود را چاپ می‌کنیم.

همانطور که مشاهده می‌شود داده‌های حذف شده تنها در حدود 1 درصد کل داده‌ها هستند.

```
[98682 rows x 9 columns]
```

شکل 5 تعداد داده‌های فیلتر شده

3. داده‌های تکراری

در فرایند جمع‌آوری داده‌ها ممکن است که ردیفی عینا تکرار بشود. حذف این داده‌ها از اهمیت ویژه‌ای برخوردار است. همانطور که مشاهده می‌شود در این دیتاست هیچ مقدار تکراری وجود نداشت.

```
Duplicate Rows:
 Empty DataFrame
```

شکل 6 تعداد داده‌های تکراری

4. عدم سازگاری داده‌ها

فرایند یادگیری مدل تنها به زبان کامپیوتری انجام می‌شود. بنابراین هر مدلی نیاز دارد که ابتدا تمام مقادیر به عدد تبدیل بشوند. در فرایندهایی مانند پردازش زبان طبیعی نیز مشابه همین عمل تکرار می‌شود.

این فرایند One Hot Encoding نیز نام دارد.

```
Unique values in 'gender':
['unknown' 'Male' 'Female' 'Other']
Unique values in 'smoking_history':
['never' 'current' 'No Info' 'former' 'ever' 'not current']
```

شکل 7 داده‌های منحصر به فرد

ابتدا مقادیر منحصر به فرد را در ستون‌هایی که عددی نیستند به دست آورده و به هرکدام یک لیبل اختصاص داده می‌شود. در نهایت این لیبل‌ها جایگزین مقادیر قبلی خواهند شد.

	gender	smoking_history
0	3	0
4	0	3
5	1	0
6	1	0
7	1	5

شکل 8 One hot Encoding

5. تعارضات و خطاهای منطقی

در این مرحله نیز خطاهایی مانند بخش مقادیر خارج از محدوده اتفاق می‌افتند. با تعریف میزان منطقی و فیلتر کردن داده‌ها این مقادیر را جایگزین می‌کنیم.

```
Invalid age data:
       gender   age  hypertension  heart_disease  smoking_history        bmi  \
226         0 -2.00             0              0                5  24.869598
271         0 -3.00             0              0                5  15.946293
296         1 -3.00             0              0                5  15.144665
360         0 -3.00             0              0                5  15.180772
458         0 -1.00             0              0                5  17.511669
...       ...   ...           ...            ...              ...        ...
99808       1 -2.00             0              0                0  28.327815
99906       0 -1.00             0              0                5  12.530383
99916       1 -1.52             0              0                5  26.196254
99940       0 -2.00             0              0                5  18.882412
99996       1 -1.00             0              0                5  18.624383

       HbA1c_level  blood_glucose_level  diabetes
226            5.7                   85         0
271            6.0                  140         0
296            5.0                  159         0
360            6.5                  130         0
458            6.0                  130         0
...            ...                  ...       ...
99808          3.5                   90         0
99906          6.6                  145         0
99916          4.5                  140         0
99940          6.1                  145         0
99996          6.5                  100         0

[1540 rows x 9 columns]
```

شکل 9 داده‌های با خطای منطقی سن

در مجموع سن در 1540 سطر خارج از محدوده 0 تا 120 است و به علت اینکه ویژگی سن مانند میزان قند خون تاثیر کاملا مستقیمی بر دیابت می‌تواند نداشته باشد، می‌توانیم آن را با مقدار میانگین 41.8 سال جایگزین می‌کنیم.

در نهایت دیتای پیش‌پردازش شده را ذخیره کرده و برای مراحل بعد استفاده خواهیم کرد.

	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
0	3	78.000000	0	1	0	101.665015	6.6	140	0
4	0	75.000000	1	1	3	23.212811	4.8	155	0
5	1	16.000000	0	0	0	28.156373	6.6	85	0
6	1	46.000000	0	0	0	16.546449	6.5	200	1
7	1	82.000000	0	0	5	25.621355	5.7	85	0
...	...	...	...	...	...	...	...	...	...
99995	1	81.000000	0	0	5	27.505580	6.2	90	0
99996	1	41.819629	0	0	5	18.624383	6.5	100	0
99997	0	70.000000	0	0	2	27.804892	5.7	155	0
99998	1	23.000000	0	0	0	35.913652	4.0	100	0
99999	1	54.000000	0	0	3	21.607675	6.6	90	0

98682 rows × 9 columns

شکل 10 داده‌های پیش‌پردازش شده

نرمال‌سازی داده‌ها در این بخش نیز می‌توانند انجام شوند اما به دلیل اینکه در بخش بعدی نیاز به نمایش تفاوت یادگیری مدل در حالت نرمال و غیر نرمال انجام شده است، این مرحله به قسمت بعدی منتقل شده است.

پیاده‌سازی و مقایسه الگوریتم‌های طبقه‌بندی

طبقه‌بندی برای پیش‌بینی نتایج و دسته‌بندی داده‌ها بر اساس الگوهای قبلی کاربرد دارد.

در این بخش دو مدل Decision Tree Classifier و SVM به دلیل توانایی مدیریت چندکلاسی و عملکرد خوب در مقابل نویز و داده‌های نامتعادل، انتخاب شدند.

آماده‌سازی داده‌ها

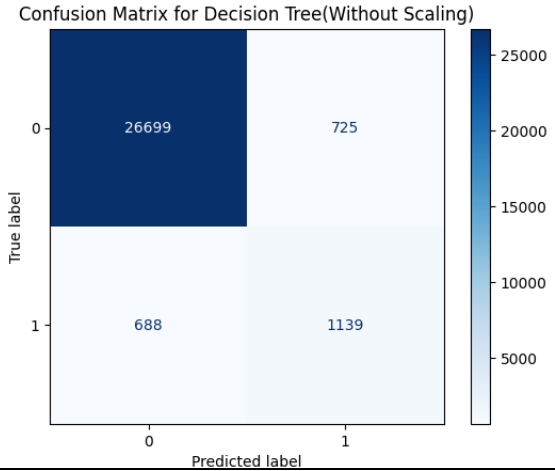
ابتدا باید داده‌ها را خوانده و ویژگی‌های مدل و خروجی را جدا کنیم. پس از آن داده‌ها به نسبت 70 به 30 برای یادگیری و تست تقسیم می‌شوند. ویژگی‌های یادگیری شامل جنسیت، سن، فشار خون، سابقه بیماری قلبی، سابقه استعمال سیگار، شاخص توده‌ی بدنی، هموگلوبین گلیکوزیله و مقدار قند خون هستند. خروجی نیز دیابتی یا عدم دیابتی بودن فرد می‌باشد.

```
Length Train Data=68252
Length Test Data=29251
```

شکل 11 تعداد داده‌های آموزش و تست

طبقه‌بندی درخت تصمیم‌گیری

در این مرحله طبقه‌بندی با دو داده‌ی نرمال و غیر نرمال انجام شد.



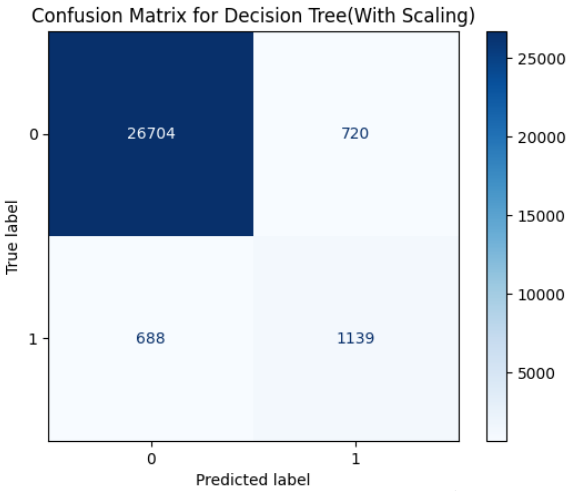

```
Classification Report for Decision Tree(Without Scaling):

              precision    recall  f1-score   support

Non-Diabetes       0.97      0.97      0.97     27424
    Diabetes       0.61      0.62      0.62      1827

    accuracy                           0.95     29251
   macro avg       0.79      0.80      0.80     29251
weighted avg       0.95      0.95      0.95     29251
```

شکل 12 ماتریس درهم‌ریختگی و معیارهای درخت تصمیم‌گیری(داده‌ی غیر نرمال)



Classification Report for Decision Tree(With Scaling):

	precision	recall	f1-score	support
Non-Diabetes	0.97	0.97	0.97	27424
Diabetes	0.61	0.62	0.62	1827
accuracy			0.95	29251
macro avg	0.79	0.80	0.80	29251
weighted avg	0.95	0.95	0.95	29251

شکل 13ماتریس درهم‌ریختگی و معیارهای درخت تصمیم‌گیری(داده‌ی نرمال)

همانطور که مشاهده می‌شود نتیجه‌ی الگوریتم درخت تصمیم‌گیری با داده‌های نرمال و غیر نرمال تفاوت چندانی با یکدیگر ندارند.

نتایج ارزیابی مدل درخت تصمیم با مقیاس‌دهی (Scaling) به شرح زیر است:

: Non-Diabetes

- **دقت (Precision):** 0.97 این به این معناست که از تمام پیش‌بینی‌های مدل مبنی بر این که فرد دیابت ندارد، 97% درست بوده‌اند.
- **حساسیت (Recall):** 0.97 مدل 97% از افرادی که واقعاً دیابت ندارند را به درستی شناسایی کرده است.
- **امتیاز F1 (F1-score):** 0.97 امتیاز F1 که ترکیبی از دقت و حساسیت است، نشان‌دهنده عملکرد عالی مدل در شناسایی افرادی است که دیابت ندارند.

:Diabetes

- **دقت (Precision):** 0.61 این به این معناست که از تمام پیش‌بینی‌های مدل مبنی بر این که فرد دیابت دارد، 61% درست بوده‌اند.
- **حساسیت (Recall):** 0.62 مدل 62% از افرادی که واقعاً دیابت دارند را به درستی شناسایی کرده است.
- **امتیاز F1 (F1-score):** 0.62 امتیاز F1 نشان می‌دهد که مدل در شناسایی موارد مثبت (افرادی که دیابت دارند) نسبت به موارد منفی (افرادی که دیابت ندارند) عملکرد کمتری داشته است.

کل داده‌هاAccuracy

- **دقت کل (Accuracy):** 0.95 این نشان می‌دهد که 95% از پیش‌بینی‌های مدل صحیح بوده است.

مقادیر میانگین:

- **Macro avg:**

- **دقت**0.79 :
- **حساسیت**0.80 :
- **امتیاز**0.80 **F1:**

این مقادیر میانگین برای تمام کلاس‌ها (دیابت و غیر دیابت) هستند و نشان می‌دهند که مدل در کل عملکرد متعادلی دارد.

Weighted avg:

- **دقت**0.95
- **حساسیت**0.95
- **F1** 0.95

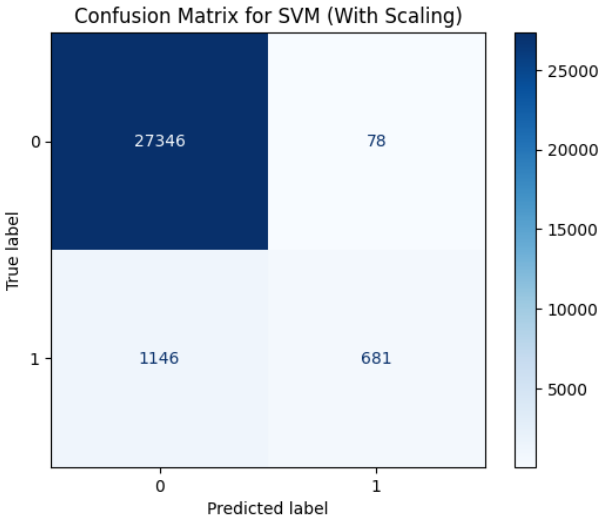
این مقادیر بر اساس تعداد نمونه‌ها در هر کلاس وزن‌دهی شده‌اند و نشان می‌دهند که مدل در شناسایی افرادی که دیابت ندارند عملکرد بسیار خوبی دارد.

تحلیل:

مدل در شناسایی افرادی که دیابت ندارند، عملکرد بسیار خوبی دارد (دقت و حساسیت بالا)، اما در شناسایی افرادی که دیابت دارند، عملکرد ضعیف‌تری دارد. دقت و حساسیت پایین‌تر برای تشخیص دیابت به این معناست که مدل در شناسایی دقیق بیماران دیابتی دچار اشتباهات زیادی است.

مدل به‌خوبی می‌تواند افراد غیر دیابتی را شناسایی کند، زیرا تعداد این افراد در داده‌ها بسیار بیشتر از افرادی است که دیابت دارند. این مسئله منجر به بالاتر بودن دقت کلی (Accuracy) می‌شود، اما برای شناسایی دقیق‌تر دیابت، ممکن است نیاز به بهینه‌سازی مدل باشد، مانند استفاده از مدل‌های پیچیده‌تر یا استفاده از تکنیک‌های مانند **over-sampling** برای داده‌های دیابتی یا تغییر وزن کلاس‌ها برای تمرکز بیشتر بر روی تشخیص افراد دیابتی.

طبقه‌بندی ماشین بردار پشتیبان



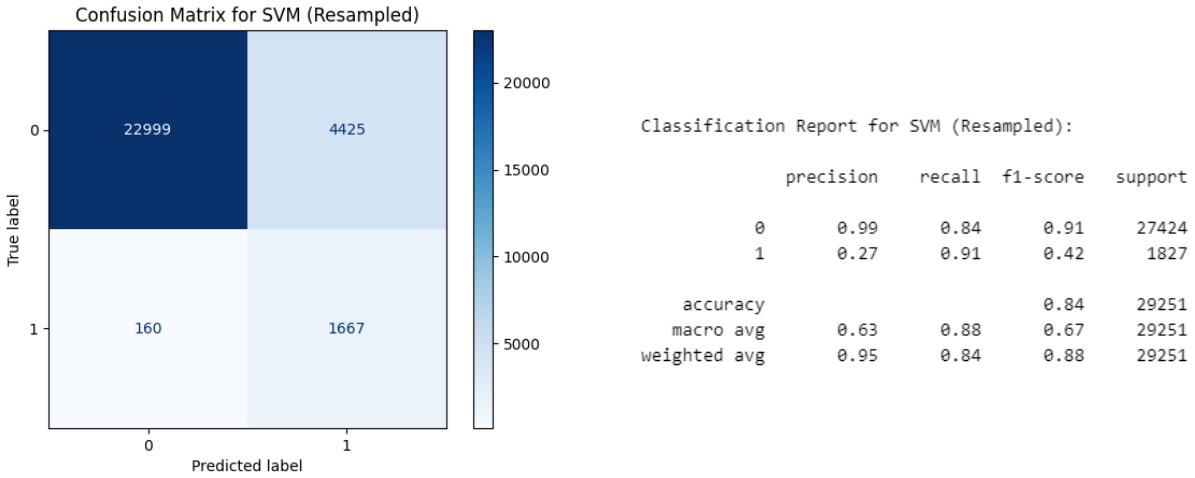
Classification Report for SVM (With Scaling):

	precision	recall	f1-score	support
0	0.96	1.00	0.98	27424
1	0.90	0.37	0.53	1827
accuracy			0.96	29251
macro avg	0.93	0.68	0.75	29251
weighted avg	0.96	0.96	0.95	29251

شکل 14ماتریس درهم‌ریختگی و مشابه‌بردار پشتیبان(داده‌ی نرمال)

همانطور که مشاهده می‌شود تعداد FP در ماتریس

بهم‌ریختگی دارای تعداد زیادی داده است. در بخش مقایسه به این قسمت پرداخته خواهد شد. اما هنگامی که داده‌ها کلاس‌ها متوازن نباشند می‌توانیم از روش‌های متوازن کردن داده‌ها استفاده کنیم. در روش SMOTE، به‌جای کپی کردن نمونه‌های موجود در کلاس‌های کمتر نماینده (کلاس اقلیت)، نمونه‌های جدیدی به‌صورت مصنوعی تولید می‌شود. این نمونه‌های جدید با استفاده از میانگین فاصله نمونه‌های موجود در کلاس اقلیت ساخته می‌شوند.



Classification Report for SVM (Resampled):

	precision	recall	f1-score	support
0	0.99	0.84	0.91	27424
1	0.27	0.91	0.42	1827
accuracy			0.84	29251
macro avg	0.63	0.88	0.67	29251
weighted avg	0.95	0.84	0.88	29251

شکل 15 ماتریس درهم‌ریختگی و مشابه‌بردار پشتیبان(داده‌ی متوازن)

همانطور که مشاهده می‌شود، عدد FP کاهش و به TN افزوده شده است. در فرایندهای طبقه‌بندی همواره یک مصالحه بین این دو مقدار وجود دارد. در داده‌های پزشکی همواره کفه‌ی ترازو را به سمتی قرار می‌دهیم که خطای نتیجه شده در بخش FP کم بشود زیرا هزینه‌ی این بخش برای افراد کمتر خواهد بود. اگر فردی بیمار باشد و به اشتباه سالم تشخیص داده بشود، با احتمال صدمه‌ی زیادی رو به رو خواهد شد.

نتایج ارزیابی مدل) **SVMبا نمونه‌برداری مجدد (**به شرح زیر است:

1.کلاس 0:(Non-Diabetes)

- **دقت (Precision):** 0.99 این به این معناست که از تمام پیش‌بینی‌های مدل مبنی بر این که فرد دیابت ندارد، 99% درست بوده‌اند.
- **حساسیت (Recall):** 0.84 مدل 84% از افرادی که واقعاً دیابت ندارند را به درستی شناسایی کرده است.
- **امتیاز F1 (F1-score):** 0.91 این امتیاز نشان‌دهنده عملکرد بسیار خوب مدل در شناسایی افراد غیر دیابتی است.

2.کلاس 1:(Diabetes)

- **دقت (Precision):** 0.27 این به این معناست که از تمام پیش‌بینی‌های مدل مبنی بر این که فرد دیابت دارد، تنها 27% درست بوده‌اند.
- **حساسیت (Recall):** 0.91 مدل 91% از افرادی که واقعاً دیابت دارند را به درستی شناسایی کرده است.
- **امتیاز F1 (F1-score):** 0.42 مدل در شناسایی افراد دیابتی عملکرد ضعیفی دارد، که به دلیل دقت پایین است، هرچند حساسیت بالا است.

3.کل داده‌ها:(Accuracy)

- **دقت کل (Accuracy):** 0.84 این نشان می‌دهد که 84% از پیش‌بینی‌های مدل صحیح بوده است، که در مقایسه با مدل درخت تصمیم (دقت 95%) پایین‌تر است.

4.مقادیر میانگین:

- **Macro avg:**
  - **دقت:** 0.63
  - **حساسیت:** 0.88
  - **F1: امتیاز** 0.67

این مقادیر میانگین برای تمام کلاس‌ها (دیابت و غیر دیابت) هستند و نشان می‌دهند که مدل در شناسایی موارد منفی (افرادی که دیابت ندارند) بسیار خوب است اما در شناسایی موارد مثبت (افرادی که دیابت دارند) عملکرد ضعیفی دارد.

- **Weighted avg:**
  - **دقت:** 0.95
  - **حساسیت:** 0.84
  - **F1: امتیاز** 0.88

این مقادیر نشان می‌دهند که مدل در شناسایی موارد منفی (افرادی که دیابت ندارند) بسیار دقیق است، اما حساسیت آن برای شناسایی موارد مثبت (افرادی که دیابت دارند) نسبتاً پایین است.

تحلیل:

- مدل **SVM**با نمونه‌برداری مجدد (Resampling) قادر است به خوبی افراد غیر دیابتی را شناسایی کند (حساسیت 0.84 و دقت 0.99)، اما در شناسایی افراد دیابتی ضعیف عمل می‌کند (دقت 0.27 و حساسیت 0.91).
- نتیجه‌ی ضعیف در دقت مدل برای کلاس دیابت نشان‌دهنده‌ی این است که مدل تعداد زیادی از پیش‌بینی‌های مثبت را اشتباه می‌زند.
- استفاده از **نمونه‌برداری مجدد** احتمالاً باعث شده است که مدل در شناسایی افراد دیابتی حساس‌تر شود، اما این حساسیت بالا بدون دقت کافی منجر به پیش‌بینی‌های اشتباه زیاد شده است.

مشکل اصلی این است که تعداد افراد دیابتی در داده‌ها بسیار کمتر از افراد غیر دیابتی است. در حالی که حساسیت مدل برای کلاس دیابت بالا است، دقت آن پایین است زیرا مدل بیشتر اوقات اشتباهاً افراد غیر دیابتی را به عنوان دیابتی شناسایی می‌کند.

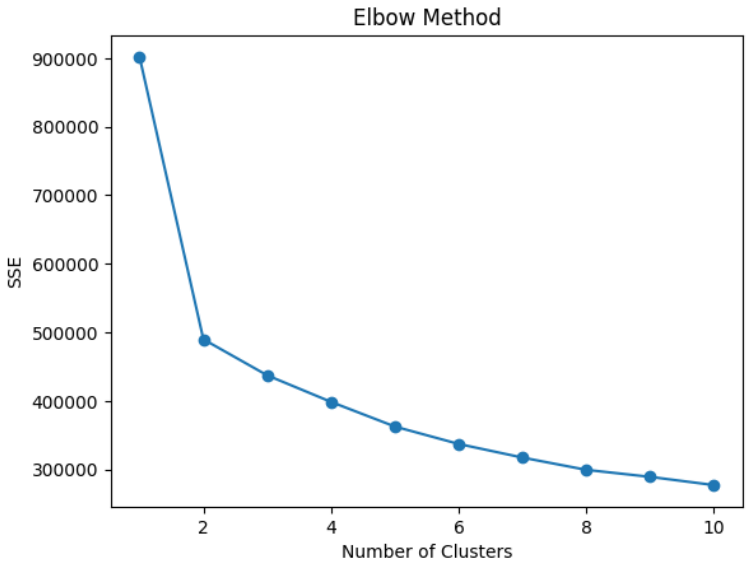
مقایسه

مدل SVM با داده‌های متوازن مدل مناسب‌تری جهت پیش‌بینی است. هنگامی که وزن داده‌های FP اهمیت دارد از معیار recall استفاده می‌شود. درحالی که در دو معیار دیگر این مدل عملکرد ضعیفی داشته است اما می‌بایست به مقدار کمینه در FP و بیشینه در TN نگاه کرد.

انجام تحلیل خوشه‌بندی

تفاوت خوشه‌بندی و طبقه‌بندی در این است که در خوشه‌بندی لیبل داده‌ها مشخص نبوده و مدل، داده‌های مشابه را در یک گروه قرار می‌دهد.

اولین مرحله در عملیات خوشه‌بندی، انتخاب تعداد خوشه‌های مناسب است. فرض میکنیم که کاملا نسبت به تعداد لیبل‌های داده بی‌اطلاع هستیم. برای انتخاب تعداد خوشه‌ها بهینه، از روش **Elbow Method** استفاده می‌کنیم که با محاسبه خطای مربعات میانگین (SSE) برای هر تعداد خوشه، تعداد خوشه‌ها را انتخاب می‌کند که در آن کاهش SSE به طور قابل توجهی کند می‌شود.



شکل 16 خروجی Elbow Method

تعداد خوشه‌ی بهینه نقطه‌ای است که قدرت بیشتری داشته باشد. مشابه با حالت دست، نقطه‌ی مشابه با آرنج را باید در نظر گرفت، که در اینجا برابر عدد 2 است.

K-means

این روش خوشه‌بندی یکی از متداول‌ترین و ساده‌ترین روش‌ها خوشه‌بندی می‌باشد. پس از آن که تعداد خوشه‌ی بهینه تعیین شد، ابتدا مرکز خوشه‌ها تعیین شده و پس از آن فاصله‌ی داده‌ها با مرکز خوشه محاسبه شده و هر داده در نزدیک‌ترین مرکز قرار می‌گیرد. خروجی خوشه‌بندی مطابق تصویر زیر است.

	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
KMeans_Cluster									
0	0.550357	-0.203034	0.048188	0.032255	4.544272	-0.159572	-0.027137	-0.036405	0.049627
1	0.622095	0.197370	0.096045	0.043406	0.491537	0.155121	0.026380	0.035390	0.095346

شکل 17 خروجی خوشه‌بندی K-means

مقادیر منفی در خوشه‌بندی نشان‌دهنده‌ی آن است که مقادیر این قسمت‌ها از میانگین کمتر بوده است. بنابراین افراد در خوشه‌ی 0 دارای سن، وزن، سطح قند خون پایین‌تر از میانگین بوده‌اند. به طور کلی قوی‌ترین ویژگی را می‌توان به افراد با جنسیت مرد، و زندگی سالم‌تری در نظر گرفت که سابقه مصرف سیگار بیشتری دارند. همچنین این افراد کمتر به دیابت مبتلا هستند.

با استفاده از تحلیلات آماری نیز می‌توان به نتیجه‌ی قبلی رسید. میزان مصرف سیگار در مردان بیشتر از زنان است و همچنین ابتلای به دیابت در سنین پایین‌تر نیز دارای احتمال کمتری است.

تحلیل دقیقتر:

1. **Cluster 0**
   - **ویژگی‌ها**:
     - **جنسیت** :مقدار مثبت نشان‌دهنده تمایل بیشتر به جنسیت خاص است (اینکه آیا این خوشه شامل بیشتر افراد مرد یا زن است).
     - **سن** :مقدار منفی نشان‌دهنده افراد مسن‌تر در این خوشه است.
     - **هیپرتنشن (فشار خون بالا)** :مقدار مثبت نشان‌دهنده افرادی است که احتمالاً فشار خون بالا دارند.
     - **بیماری قلبی** :مقدار مثبت، ممکن است نشان‌دهنده افرادی با بیماری‌های قلبی باشد.
     - **سابقه سیگار کشیدن** :مقدار مثبت نشان‌دهنده افراد با سابقه سیگار کشیدن است.
     - **BMI**:مقدار بالا نشان‌دهنده افرادی با BMI بالا است.
     - **سطح HbA1c**: مقدار منفی که نشان‌دهنده کنترل نسبی قند خون است.
     - **سطح قند خون** :مقدار منفی که نشان‌دهنده سطح قند خون نسبی کنترل‌شده است.
     - **دیابت** :مقدار کم به این معنی است که احتمال دیابت کمتر است.
   - **نتیجه‌گیری** :این خوشه احتمالاً شامل افرادی با BMI بالا، فشار خون بالا و بیماری‌های قلبی است که از نظر قند خون نیز مشکلات کمتری دارند.
2. **Cluster1**
   - **ویژگی‌ها**:
     - **جنسیت** :مقدار مثبت نشان‌دهنده تمایل به جنسیت خاص است.
     - **سن** :مقدار مثبت، احتمالاً نشان‌دهنده افراد جوانتر است.
     - **هیپرتنشن (فشار خون بالا)** :مقدار مثبت، به این معنی که این افراد بیشتر فشار خون بالا دارند.
     - **بیماری قلبی** :مقدار مثبت، ممکن است نشان‌دهنده افرادی با مشکلات قلبی باشد.
     - **سابقه سیگار کشیدن** :مقدار مثبت، افراد با سابقه سیگار کشیدن را نشان می‌دهد.
     - **BMI**:مقدار پایین‌تر نشان‌دهنده افرادی با BMI متوسط یا پایین است.
     - **سطح HbA1c**: مقدار مثبت که نشان‌دهنده مشکلات کنترل قند خون است.
     - **سطح قند خون** :مقدار مثبت، که نشان‌دهنده سطح قند خون بالاتر است.
     - **دیابت** :مقدار متوسط به این معنی است که احتمال دیابت بیشتر است.

- **نتیجه‌گیری** :این خوشه شامل افرادی است که بیشتر به دیابت و مشکلات قند خون مبتلا هستند، همراه با فشار خون بالا و بیماری‌های قلبی. افراد در این خوشه ممکن است سبک زندگی ناسالمی مانند سیگار کشیدن داشته باشند.

مقایسه نتایج:

در این خوشه‌ها، تفاوت اصلی بین **BMI**، **سطح قند خون** و **ریسک بیماری‌های قلبی** است. خوشه اول شامل افرادی است که به نظر می‌رسد دارای BMI بالا و سطح قند خون کنترل‌شده‌تری باشند، در حالی که خوشه دوم شامل افرادی است که احتمالاً بیشتر با دیابت و فشار خون بالا دست و پنجه نرم می‌کنند.

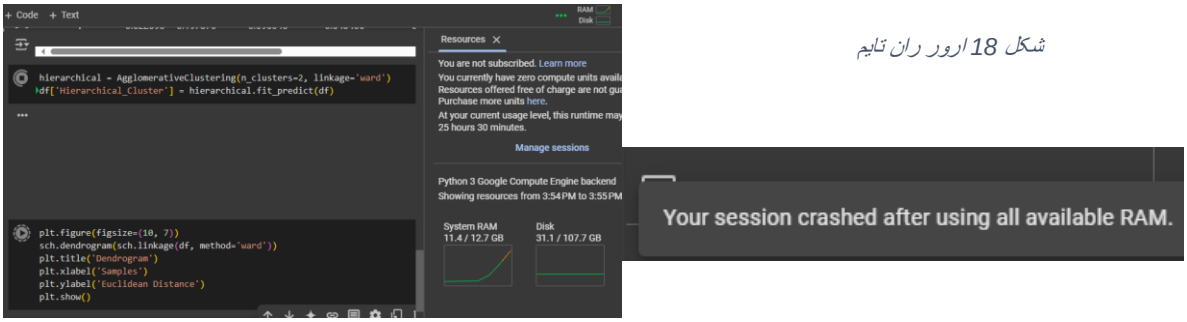
الگوریتم سلسله مراتبی

الگوریتم‌های سلسله مراتبی دارای عمق و پیچیدگی بیشتری هستند و بنابراین به سخت‌افزار قوی‌تری برای تحلیل نیاز الگوریتم های خوشه بندی سلسله مراتبی به دو دسته زیر تقسیم می شوند دارند.

- الگوریتم های سلسله مراتبی تجمعی – در الگوریتم های سلسله مراتبی تجمعی ، هر نقطه داده به عنوان یک خوشه واحد در نظر گرفته می شود و سپس بطور متوالی جفت خوشه ها را ادغام یا جمع می کند (رویکرد از پایین به بالا). سلسله مراتب خوشه ها به عنوان یک دندروگرام یا ساختار درختی نشان داده می شود.

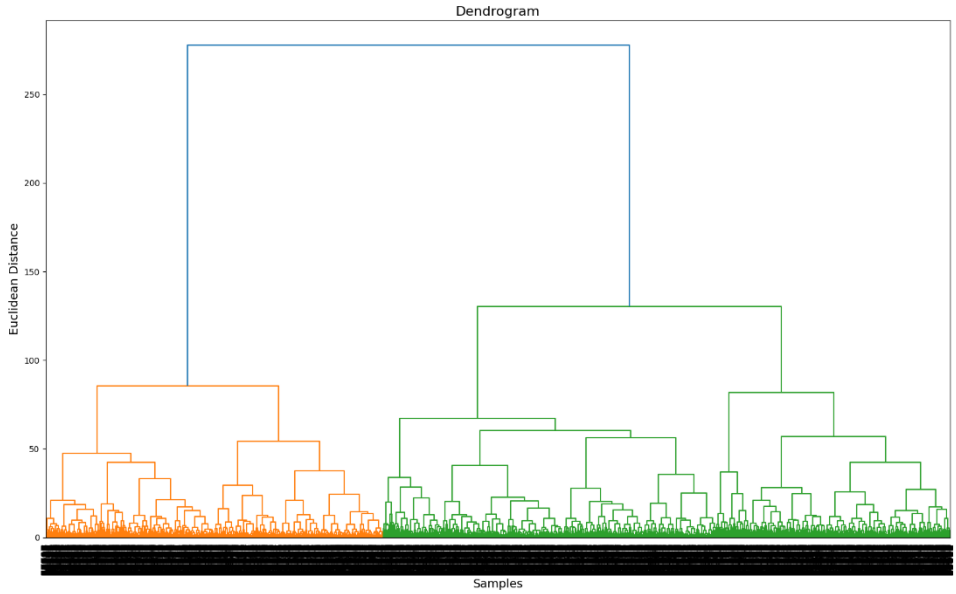
- الگوریتم های سلسله مراتبی تقسیم کننده – از سوی دیگر ، در الگوریتم های سلسله مراتبی تقسیم پذیر ، تمام نقاط داده به عنوان یک خوشه بزرگ در نظر گرفته می شوند و روند خوشه بندی شامل تقسیم (رویکرد از بالا به پایین) خوشه بزرگ به خوشه های کوچک مختلف است.

کد این بخش نوشته شد و در کولب ران شد، اما به علت محدودیت فضای رم، ران تایم کرش کرده و مدل آموزش داده نشد.



شکل 18 ارور ران تایم

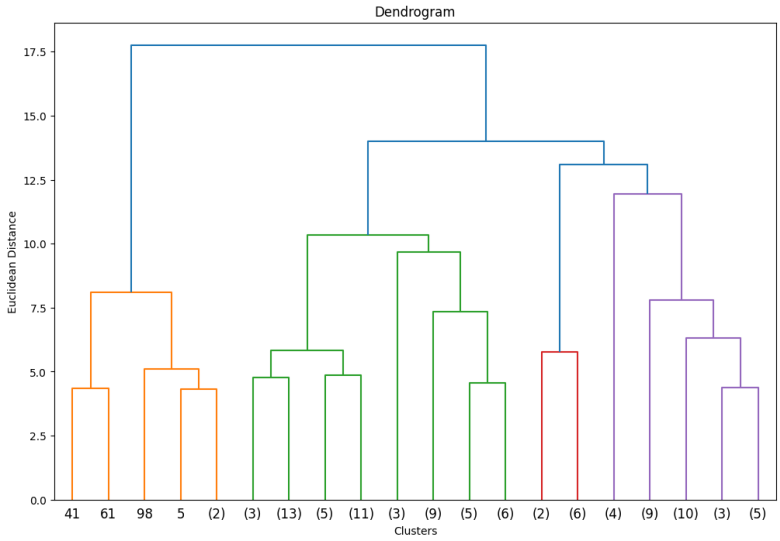
پس از روش سمپلینگ استفاده میکنیم. چون کل داده های ما زیاد است، هر کتابخانه از خوشه بندی که من استفاده کردم رم کرش شد. از 1000 نمونه استفاده میکنم. نتیجه ای که دریافت کردم:



تحلیل دندروگرام:

1. **فاصله میان خوشه‌ها**:
   - محور عمودی (Euclidean Distance) نشان‌دهنده میزان فاصله یا تفاوت بین خوشه‌هاست.
   - هر چه یک خط (پیوند خوشه‌ها) بالاتر باشد، نشان می‌دهد که این خوشه‌ها تفاوت بیشتری دارند.
   - در اینجا، دو خوشه اصلی در فاصله‌ای بسیار زیاد (بیش از 200) به هم متصل شده‌اند که نشان‌دهنده تفاوت زیاد بین این دو گروه اصلی است.
2. **تعداد خوشه‌ها**:
   - اگر به برش‌هایی در سطوح مختلف محور عمودی نگاه کنیم:
     - در فاصله تقریباً 50، چندین خوشه کوچک‌تر وجود دارد.
     - در فاصله بالاتر (100)، خوشه‌ها به دو دسته اصلی تقسیم شده‌اند.
   - بنابراین، به نظر می‌رسد دو خوشه اصلی وجود دارد که خود شامل زیرخوشه‌هایی هستند.
3. **تراکم نمونه‌ها**:
   - تعداد بسیار زیاد نمونه‌ها در محور افقی باعث شده که برچسب‌های نمونه‌ها خوانا نباشند. این نشان می‌دهد که حجم داده زیاد است.
   - برای بررسی دقیق‌تر، بهتر است داده‌ها به زیرمجموعه‌های کوچکتر تقسیم شوند.

داده هارا کاهش دادم به 100 تا:



خروجی خوشه‌بندی ما شامل دو خوشه Cluster 0وCluster 1 است. تحلیل این خروجی به شرح زیر است:

1. (Silhouette Score):

- بهترین امتیاز سیلوئت برای 2 خوشه به دست آمده است (0.45). این نشان می‌دهد که داده‌ها به خوبی به دو گروه تقسیم شده‌اند.
- با افزایش تعداد خوشه‌ها (3، 4، و 5)، امتیاز کاهش پیدا کرده است که نشان می‌دهد خوشه‌بندی با تعداد بیشتر خوشه‌ها کیفیت پایین‌تری دارد.

تحلیل خوشه‌ها:

خوشه 0:(Cluster 0)

- **gender جنسیت:** مقدار 0.648 نشان‌دهنده این است که بیشتر داده‌های این خوشه مربوط به جنسیتی با مقدار کدگذاری نزدیک به 0.65 است.
- **Age سن :** مقدار نزدیک به 0 نشان می‌دهد که افراد این خوشه در سنین نزدیک به میانگین داده‌ها قرار دارند.
- **hypertension فشار خون بالا:** مقدار 0.085 نشان‌دهنده این است که تنها 8.5٪ افراد این خوشه دچار فشار خون بالا هستند.
- **heart_disease بیماری قلبی :**مقدار 0.042 (حدود 4٪) نشان می‌دهد که درصد کمی از افراد این خوشه بیماری قلبی دارند.
- **smoking_history تاریخچه سیگار کشیدن:** مقدار نزدیک به 2.36 نشان‌دهنده تاریخچه سیگار کشیدن متوسط است.
- **Bmi شاخص توده بدنی:** مقدار بسیار نزدیک به 0 نشان می‌دهد که BMI افراد این خوشه در حدود میانگین کل داده‌ها قرار دارد.

- **HbA1c_level سطح HbA1c** : مقدار -0.09 نشان‌دهنده سطح HbA1c کمتر از میانگین است.
- **blood_glucose_level سطح قند خون:** مقدار -0.29 نشان‌دهنده سطح قند خون کمتر از میانگین است.
- **diabetes دیابت:** مقدار 0 نشان‌دهنده این است که بیشتر افراد این خوشه دیابت ندارند.

خوشه Cluster1

- **gender جنسیت:** مقدار 1.0 نشان‌دهنده این است که جنسیت تمام افراد این خوشه یکسان است.
- **age سن:** مقدار 1.09 نشان می‌دهد که افراد این خوشه به طور میانگین سن بالاتری دارند.
- **Hypertension فشار خون بالا:** مقدار 0.33 نشان‌دهنده این است که حدود 33٪ از افراد این خوشه فشار خون بالا دارند.
- **heart_disease بیماری قلبی:** مقدار 0.33 نشان‌دهنده این است که 33٪ از افراد این خوشه بیماری قلبی دارند.
- **smoking_history تاریخچه سیگار کشیدن:** مقدار 2.66 نشان‌دهنده تاریخچه سیگار کشیدن کمی بالاتر از میانگین است.
- **bmi شاخص توده بدنی:** مقدار -0.23 نشان‌دهنده BMI کمتر از میانگین است.
- **HbA1c_level سطحHbA1c** : مقدار 1.24 نشان‌دهنده سطح HbA1c بالاتر از میانگین است.
- **blood_glucose_level سطح قند خون:** مقدار 1.90 نشان‌دهنده سطح قند خون بسیار بالاتر از میانگین است.
- **diabetes دیابت:** مقدار 1 نشان می‌دهد که تمام افراد این خوشه دیابت دارند.

- **خوشه 0 :**شامل افرادی است که وضعیت سلامتی بهتری دارند. سطح قند خون و HbA1c در این خوشه کمتر از میانگین است. همچنین فشار خون بالا و بیماری قلبی در این خوشه به ندرت دیده می‌شود.
- **خوشه 1 :**شامل افرادی است که وضعیت سلامتی ضعیف‌تری دارند. این افراد سن بالاتری دارند و درصد بالایی از آن‌ها دیابت، فشار خون بالا و بیماری قلبی دارند. همچنین سطح HbA1c و قند خون آن‌ها به طور معناداری بالاتر از میانگین است.

می‌توانیم مثلا برای هر خوشه راهبردهای بهداشتی مناسب را طراحی کنیم .

- برای خوشه 0: تمرکز بر پیشگیری از مشکلات سلامتی آینده.
- برای خوشه 1: ارائه مراقبت‌های خاص برای دیابت و فشار خون بالا.

مقایسه ی الگوریتم k-means و سلسله مراتبی:

1. **شباهت‌ها**:
   - هر دو الگوریتم خوشه‌بندی (K-Means و Hierarchical Clustering) خوشه‌هایی را ایجاد کرده‌اند که به وضوح نماینده گروه‌هایی با ویژگی‌های مشابه هستند.

- در هر دو الگوریتم، خوشه‌ها به دو گروه اصلی تقسیم می‌شوند که به‌طور قابل توجهی از نظر ویژگی‌هایی مانند **BMI**، **سطح قند خون**، **دیابت** و **سابقه سیگار کشیدن** متفاوت هستند.

2. **تفاوت‌ها**:

- **K-Means**در خوشه‌ها مقدار **BMI**بالاتری در خوشه 0 نشان می‌دهد، در حالی که **Hierarchical Clustering**مقدار پایین‌تری از **BMI**را در خوشه مشابه (خوشه 0) ثبت کرده است.
- در خوشه 1، **K-Means**سطح **قند خون** و **HbA1c**بالاتری نشان می‌دهد، در حالی که **Hierarchical Clustering**بیشتر بر روی **هیپرتنشن** و **بیماری قلبی** تأکید دارد.

3. **Silhouette Score**:

- **K-Means**معمولاً دارای **Silhouette Score**بالاتری است که نشان‌دهنده کیفیت بهتر خوشه‌ها است.
- **Hierarchical Clustering**با خوشه 2 بهینه‌تر عمل کرده است (Silhouette Score: 0.45)و بعد از آن با تعداد خوشه‌های بیشتر کیفیت تحلیل کاهش می‌یابد.

تحلیل عملکرد الگوریتم‌ها:

- **K-Means**به‌دلیل سادگی و سرعت بالا، معمولاً برای داده‌های بزرگ و با ساختار خوشه‌ای مشخص مناسب‌تر است.
- **Hierarchical Clustering**از نظر ارائه تحلیل سلسله‌مراتبی دقیق‌تر عمل می‌کند و ممکن است برای درک روابط بین خوشه‌ها مفیدتر باشد.

- اگر هدف ما **کاهش تعداد خوشه‌ها** و تحلیل ساده‌تر است، **K-Means**بهترین انتخاب است.
- اگر بخواهیم درک عمیق‌تری از ساختار داده‌ها و روابط بین خوشه‌ها داشته باشیم، **Hierarchical Clustering**ممکن است بهتر عمل کند، به‌ویژه اگر می‌خواهیم تعداد دقیق خوشه‌ها را بدون نیاز به پیش‌فرض‌های الگوریتمی تعیین کنیم.

تحلیل های دیگر:

افرادی که دارای **سن بالا** و **فشار خون بالا** هستند، به دلیل تغییرات فیزیولوژیکی، بیماری‌های مزمن و رفتارهای بهداشتی مشابه، بیشتر در یک خوشه قرار می‌گیرند. این شرایط به طور طبیعی باعث می‌شود که ویژگی‌های سلامت این افراد مشابه یکدیگر باشد و در نتیجه الگوریتم‌های خوشه‌بندی این افراد را در یک گروه قرار دهند.

باتشکر