

**دانشکده مهندسی کامپیوتر**

**گزارش پروژه درس داده کاوی**

پیاده سازی و مقایسه الگوریتم های مختلف بر روی دیتاست دیابت

**دانشجو:**

امیر حسین علی نژاد(4001231083)

**استاد:**

دانشپور، نگین

**بهمن 1403**



**چکیده**

**این گزارش به بررسی و مقایسه‌ی روش‌های مختلف یادگیری ماشین برای پیش‌بینی ریسک ابتلا به دیابت می‌پردازد. در این راستا، از چهار الگوریتم دسته‌بندی شامل درخت تصمیم، ماشین بردار پشتیبان (SVM)، نزدیک‌ترین همسایه (KNN) و رگرسیون لجستیک، و دو روش خوشه‌بندی K-means و خوشه‌بندی سلسله مراتبی استفاده شده است. عملکرد مدل‌های دسته‌بندی با معیارهای دقت، بازخوانی، دقت مثبت و امتیاز F1 ارزیابی شده و خوشه‌بندی‌ها با امتیاز سیلوئت و شاخص کالینسکی-هاراباس مورد بررسی قرار گرفته‌اند. نتایج نشان می‌دهد که در بخش دسته‌بندی، مدل‌های درخت تصمیم و KNN عملکرد بهتری دارند و در بخش خوشه‌بندی، روش K-means مناسب‌تر است. در نهایت، پیشنهاداتی برای بهبود عملکرد مدل‌ها ارائه شده است.**

**کد ها و محتوای مربوط به این پروژه ار طریق لینک زیر قابل مشاهده میباشد:**[**invaha/datamining at master**](https://github.com/invaha/datamining/tree/master)

**فهرست مطالب**

**عنوان** **صفحه**

[بخش 1 مقدمه 4](#_بخش_اول_)

[بخش 2داده های ورودی و پیش‌پردازش آنها 8](#_بخش_دوم_)

[بخش 3 الگوریتم های انتخابی و پیاده سازی 13](#_بخش_سوم_)

[بخش 4 نتیجه گیری 22](#_بخش_چهارم_)

# بخش اول مقدمه

**۱. مقدمه و هدف پروژه:**

این پروژه به منظور تحلیل و بررسی مجموعه داده‌ی cleaned\_data.csv با استفاده از تکنیک‌های داده‌کاوی، به ویژه خوشه‌بندی و طبقه‌بندی، انجام شده است. هدف اصلی، کشف الگوهای پنهان در داده‌ها، گروه‌بندی نمونه‌های مشابه و پیش‌بینی وضعیت متغیر هدف (در صورت وجود) است. این تحلیل به ما کمک می‌کند تا درک بهتری از داده‌ها پیدا کنیم، روابط بین ویژگی‌ها را شناسایی کنیم و در نهایت، تصمیمات بهتری بر اساس داده‌ها اتخاذ کنیم.

این پروژه شامل مراحل زیر است:

* **پیش‌پردازش داده‌ها:** این مرحله شامل پاکسازی داده‌ها، مدیریت مقادیر گمشده، تبدیل داده‌های دسته‌ای به عددی و مقیاس‌بندی داده‌ها است. اهمیت این مرحله در این است که داده‌های خام معمولاً دارای نویز، مقادیر گمشده و فرمت‌های ناسازگار هستند که می‌توانند باعث کاهش دقت و کارایی مدل‌ها شوند. پیش‌پردازش داده‌ها این مشکلات را برطرف می‌کند و داده‌ها را برای تحلیل آماده می‌کند.
* **خوشه‌بندی (Clustering):** در این مرحله، از الگوریتم‌های خوشه‌بندی K-means و خوشه‌بندی سلسله مراتبی برای گروه‌بندی نمونه‌های مشابه در داده‌ها استفاده می‌شود. هدف از خوشه‌بندی، پیدا کردن گروه‌هایی از نمونه‌ها است که ویژگی‌های مشابهی دارند. این کار به ما کمک می‌کند تا ساختار داده‌ها را بهتر درک کنیم و الگوهای پنهان در آن‌ها را کشف کنیم. خوشه‌بندی در مواردی که متغیر هدف وجود ندارد (یادگیری بدون نظارت) یا زمانی که می‌خواهیم گروه‌های طبیعی در داده‌ها را پیدا کنیم، کاربرد دارد.
* **طبقه‌بندی (Classification):** در این مرحله، از الگوریتم‌های طبقه‌بندی مانند درخت تصمیم، ماشین بردار پشتیبان (SVM)، k-نزدیکترین همسایه (KNN) و رگرسیون لجستیک برای پیش‌بینی مقدار متغیر هدف بر اساس سایر ویژگی‌ها استفاده می‌شود. هدف از طبقه‌بندی، ساخت مدلی است که بتواند نمونه‌های جدید را بر اساس ویژگی‌هایشان به درستی به کلاس‌های مختلف دسته‌بندی کند. طبقه‌بندی در مواردی که متغیر هدف وجود دارد و می‌خواهیم آن را پیش‌بینی کنیم (یادگیری با نظارت)، کاربرد دارد.
* **ارزیابی مدل‌ها:** در این مرحله، عملکرد مدل‌های خوشه‌بندی و طبقه‌بندی با استفاده از معیارهای مختلف ارزیابی می‌شود. این کار به ما کمک می‌کند تا بهترین مدل را برای داده‌های خود انتخاب کنیم و از عملکرد آن در داده‌های جدید اطمینان حاصل کنیم.

**اهمیت خوشه‌بندی در تحلیل داده‌ها:**

خوشه‌بندی یک تکنیک قدرتمند در تحلیل داده‌ها است که به ما امکان می‌دهد تا:

* **کشف الگوهای پنهان:** با گروه‌بندی نمونه‌های مشابه، می‌توان الگوها و روابطی را که به صورت مستقیم قابل مشاهده نیستند، کشف کرد.
* **کاهش پیچیدگی داده‌ها:** با دسته‌بندی نمونه‌ها به خوشه‌های کوچکتر، می‌توان تحلیل داده‌ها را ساده‌تر کرد.
* **بهبود عملکرد سایر مدل‌ها:** می‌توان از نتایج خوشه‌بندی به عنوان ویژگی‌های جدید در مدل‌های طبقه‌بندی یا رگرسیون استفاده کرد و عملکرد آنها را بهبود بخشید.
* **بخش‌بندی مشتریان:** در بازاریابی، می‌توان از خوشه‌بندی برای بخش‌بندی مشتریان بر اساس رفتار خرید، ویژگی‌های جمعیت‌شناختی و غیره استفاده کرد.

**اهمیت طبقه‌بندی در تحلیل داده‌ها:**

طبقه‌بندی نیز یک تکنیک مهم در تحلیل داده‌ها است که به ما امکان می‌دهد تا:

* **پیش‌بینی متغیر هدف:** با ساخت مدل‌های طبقه‌بندی، می‌توان مقدار متغیر هدف را برای نمونه‌های جدید پیش‌بینی کرد.
* **اتخاذ تصمیمات بهتر:** پیش‌بینی دقیق متغیر هدف می‌تواند به اتخاذ تصمیمات بهتر در زمینه‌های مختلف کمک کند.
* **شناسایی عوامل مؤثر:** با تحلیل مدل‌های طبقه‌بندی، می‌توان عوامل مؤثر بر متغیر هدف را شناسایی کرد.

# بخش دوم داده های ورودی و پیش‌پردازش آنها

**۲. داده‌های ورودی و پیش‌پردازش داده‌ها:**

**داده‌های ورودی:**

مجموعه داده‌ی اصلی مورد استفاده در این پروژه، modified\_diabetes\_prediction\_dataset نام دارد. این مجموعه داده شامل اطلاعات مربوط به افراد و برخی شاخص‌های پزشکی آن‌ها است. ویژگی‌های موجود در این مجموعه داده عبارتند از:

* gender (جنسیت): شامل مقادیر Male، Female و Unknown.و Other
* age (سن): سن افراد به صورت عددی.(اعداد از منفی 5 تا 84)
* hypertension (فشار خون بالا): یک ویژگی باینری (۰ و ۱) که نشان می‌دهد فرد فشار خون بالا دارد یا خیر.
* heart\_disease (بیماری قلبی): یک ویژگی باینری (۰ و ۱) که نشان می‌دهد فرد بیماری قلبی دارد یا خیر.
* smoking\_history (سابقه سیگار کشیدن): شامل دسته‌های مختلفی مانند never، No Info و غیره.( 'yes', 'current', 'never', 'No Info', 'former', 'ever','not current')
* bmi (شاخص توده بدنی): شاخص توده بدنی افراد به صورت عددی.(اعداد از 3 تا 96)
* HbA1c\_level : نشان‌دهنده میانگین سطح قند خون در طول دو تا سه ماه گذشته است.
* blood\_glucose\_level (سطح گلوکز خون): سطح گلوکز خون افراد.
* diabetes (دیابت): یک ویژگی باینری (۰ و ۱) که نشان می‌دهد فرد دیابت دارد یا خیر. این ویژگی به عنوان متغیر هدف در مدل‌های طبقه‌بندی استفاده می‌شود و در خوشه‌بندی مورد استفاده قرار نگرفته است.

**نحوه بارگذاری داده‌ها:**

داده‌ها با استفاده از کتابخانه pandas در پایتون و تابع pd.read\_csv() از فایل CSV بارگذاری شدند:

import pandas as pd

data = pd.read\_csv('modified\_diabetes\_prediction\_dataset.csv')

**پیش‌پردازش داده‌ها:**

به منظور آماده‌سازی داده‌ها برای تحلیل و بهبود عملکرد مدل‌ها، مراحل پیش‌پردازش زیر انجام شد:

1. **مدیریت داده‌های گمشده:**
   * در برخی از ویژگی‌ها، مقادیر گمشده وجود داشت. به عنوان مثال، در ستون gender مقدار unknown به عنوان داده گمشده در نظر گرفته شدند.
   * برای مدیریت این مقادیر گمشده، از روش جایگزینی با میانگین استفاده شد چون در مواردی که تصمیم به بهبود ویژگی گرفته شد(جایگزینی به جای حذف) موارد نرمال بودند و میتوانستیم از میانگین استفاده کنیم.
   * در بعضی موارد که تعداد این موراد گم شده خیلی کم بود یا به نسبت کل داده ها مقدار کمی داشت از آنها صرف نظر شد و آنهارا حذف کردیم(مثلا در ستون جنسیت 19 آیتم کلا مشخص نبود چه جنسیتی دارند و برای همین آنهارا حذف نمودیم.)
2. **تبدیل ویژگی‌های دسته‌ای:**
   * ویژگی‌های دسته‌ای gender و smoking\_history با استفاده از LabelEncoder به مقادیر عددی تبدیل شدند. این تبدیل برای استفاده از الگوریتم‌های یادگیری ماشین که با داده‌های عددی کار می‌کنند، ضروری است.
3. **مقیاس‌بندی ویژگی‌ها:**
   * از آنجایی که الگوریتم‌های خوشه‌بندی مانند K-means و برخی از الگوریتم‌های طبقه‌بندی مانند SVM به مقیاس ویژگی‌ها حساس هستند، از StandardScaler برای استانداردسازی ویژگی‌ها استفاده شد. این کار باعث می‌شود که تمام ویژگی‌ها میانگین ۰ و انحراف معیار ۱ داشته باشند.
4. **چک کردن املای یکتای کلمات در ستون های مختلف:**

* مثلا در ستون جنسیت کامل بررسی شده که املای هر جنسیت (male, female) یکتا باشند و همگی مثل هم باشند تا در تحلیل به مشکل نخوریم.

1. **تبدیل مقادیر به مقادیر معنا دار:**

* در ستون های مختلف چک شده که مقادیر موجود با معنا هستند یا نه مثلا در مورد ستون اول نباید سن منفی میداشتیم که تمامی سن های منفی به مقدار میانگین جایگزین شدند، این میانگین بعد از در نظر نگرفتن این داده های منفی محاسبه شده و تمامی موارد به مقدار زیاد در نوت بوک پیوست شده به گزارش آمده است.
* در برخی ستون ها مانند بی ام آی، از این مورد که امکان اینکه یک نفر بی ام آی 3 یا 95 داشته باشد صرف نظر شد.

1. **صرف نظر از برخی موارد که تکرار زیادی داشتند:**

* در مورد ستون مرتبط به سیگاری بودن افراد بررسی شد که هر کدام از موارد چه فراوانی ای دارند و به نظر آمد فراوانی کسانی که اطلاعاتی درباره سیگاری بودن یا نبودن آنها است زیاد است(تقریبا 35 هزار نفر) و برای همین از این افراد صرف نظر شد و کاری نکردم چون تغییر در آنها میتوانست باعث انحراف دیتاست شود و در واقعیت هم میشود ک فردی از ما پنهان کند که سیگاری است یا نه.

**ضرورت مراحل پیش‌پردازش:**

مراحل پیش‌پردازش داده‌ها برای دلایل زیر ضروری بودند:

* **بهبود کیفیت داده‌ها:** مدیریت داده‌های گمشده و اصلاح داده‌های غلط باعث می‌شود که داده‌ها دقیق‌تر و قابل اعتمادتر باشند.
* **بهبود عملکرد مدل‌ها:** تبدیل ویژگی‌های دسته‌ای به عددی و مقیاس‌بندی ویژگی‌ها باعث می‌شود که الگوریتم‌های یادگیری ماشین به درستی کار کنند و عملکرد بهتری داشته باشند.
* **جلوگیری از بایاس:** مدیریت داده‌های گمشده و مقیاس‌بندی ویژگی‌ها می‌تواند از ایجاد بایاس در مدل‌ها جلوگیری کند.
* **حفظ ارزش داده ها**: در این متد ها ما نیاز داریم که داده هایمان با ارزش باشند و مثلا سن یک فرد منفی نباشد.

# بخش سوم الگوریتم های انتخابی و پیاده سازی

**۳. الگوریتم‌های انتخابی و پیاده‌سازی:**

در این بخش، الگوریتم‌های خوشه‌بندی K-means و خوشه‌بندی سلسله مراتبی و همچنین الگوریتم‌های طبقه‌بندی رگرسیون لاجستیک، درخت تصمیم، SVM و KNN که برای تحلیل داده‌ها انتخاب شده‌اند، شرح داده می‌شوند و دلایل انتخاب آن‌ها و نتایج حاصل از پیاده‌سازی آن‌ها به همراه تحلیل‌ها ارائه می‌گردد.

**خوشه‌بندی:**

در این پروژه، از دو الگوریتم خوشه‌بندی زیر استفاده شده است:

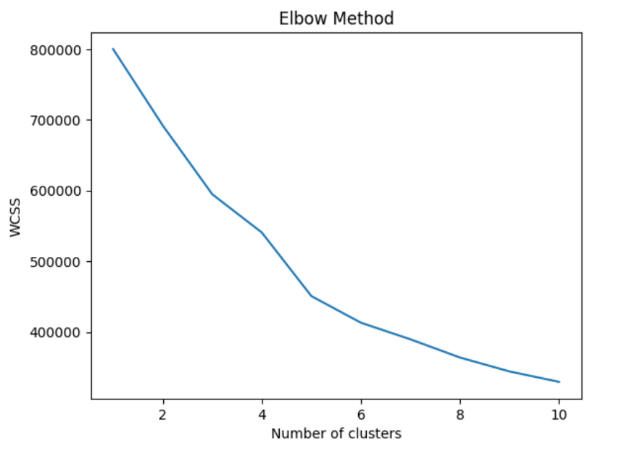
* **K-means** **:** این الگوریتم یک روش خوشه‌بندی مبتنی بر مرکز است که داده‌ها را به k خوشه تقسیم می‌کند، به طوری که هر نقطه داده به خوشه‌ای اختصاص داده می‌شود که نزدیک‌ترین میانگین (مرکز خوشه) را داشته باشد. K-means به دلیل سادگی و سرعت بالا، به خصوص برای داده‌های بزرگ، بسیار محبوب است.
* **خوشه‌بندی سلسله مراتبی: (Hierarchical Clustering)** این الگوریتم یک ساختار سلسله مراتبی از خوشه‌ها ایجاد می‌کند. این ساختار به صورت یک درخت (دندروگرام) نمایش داده می‌شود که نشان می‌دهد خوشه‌ها چگونه به تدریج با هم ادغام می‌شوند. خوشه‌بندی سلسله مراتبی برای کشف روابط سلسله مراتبی بین خوشه‌ها و همچنین زمانی که تعداد خوشه‌ها از پیش مشخص نیست، مفید است.

**دلایل انتخاب این الگوریتم‌ها:**

* K-means به دلیل سادگی، سرعت و کارایی در خوشه‌بندی داده‌های عددی انتخاب شد.
* خوشه‌بندی سلسله مراتبی برای بررسی ساختار سلسله مراتبی احتمالی در داده‌ها و همچنین برای مقایسه نتایج با K-means انتخاب شد.

**نتایج و تحلیل خوشه‌بندی K-means:**

* برای تعیین تعداد بهینه خوشه‌ها (k)، از روش آرنج (Elbow Method) و امتیاز سیلوئت (Silhouette Score) استفاده شد.

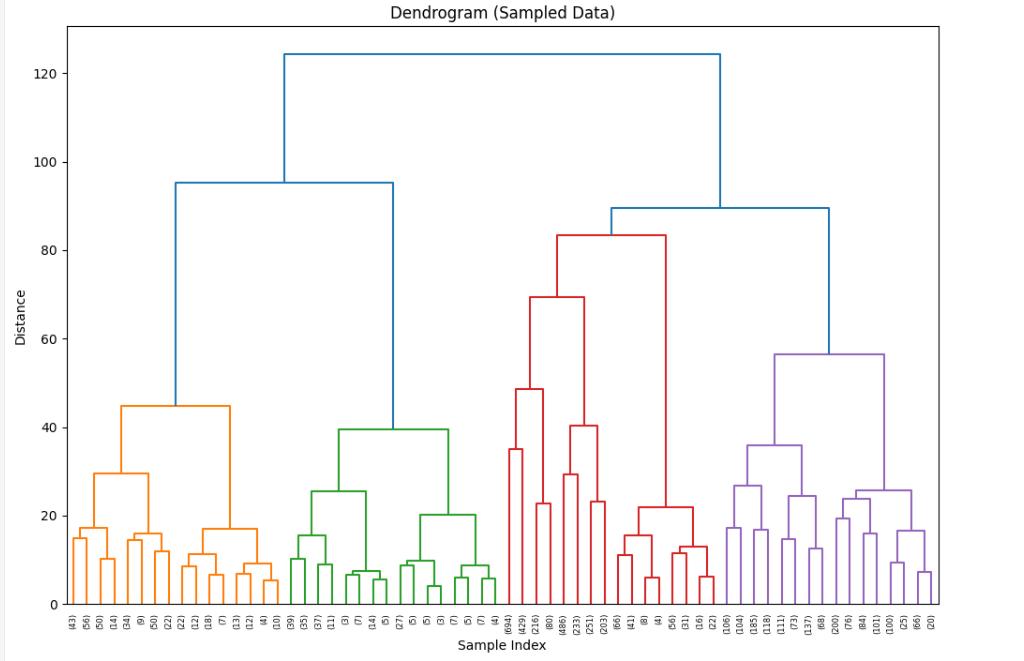


* این نمودار نشان می‌دهد که در مورد این روش رفتن از دو دسته به 3 و بیشتر موجب پایین آمدن وی سی سی شده است. که این یعنی مجموع مربعات درون خوشه‌ای کم میشود. این معیار نشان میدهد که چقدر نقاط داده در داخل خوشه‌های خود پراکنده هستند. بنابر این نمودار ما تعداد خوشه هارا 4 انتخاب میکنیم چون بعد از آن عملا تفاوت زیادی ایجاد نمی‌شود(البته که 2 نیز مورد مطلوبی خواهد بود)
* از طرفی این نمودار به ما میگوید کا را 8 یا 9 بگیریم که حالا ما همه موارد را تست میکنیم.
* در نهایت، کا مینز با مقادیر 2 تا 10 اجرا شد و نتایج در کد قابل مشاهده است. که در نهایت منجر به انتخاب بین دو تعداد 5 و 9 شد. که با رسم نمودار های پراکندگی به این نتیجه رسیدیم که در نهایت خوشه با کا 5 به دلیل تفسیرپذیری بیشتر، جلوگیری از بیش‌برازش و ارائه خوشه‌های معنادارتر، انتخاب بهتری برای این داده‌ها است. این موضوع نشان می‌دهد که صرفاً اتکا به معیارهای کمی مانند امتیاز سیلوئت کافی نیست و باید تحلیل کیفی خوشه‌ها و هدف پروژه نیز در نظر گرفته شود.
* امتیاز سیلوئت برای این خوشه‌بندی برابر با 0.197و شاخص Calinski-Harabasz برابر با 19326.916بود.

**تحلیل خوشه‌ها (K-means):**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| خوشه | جنسیت | سن | فشار خون | بیماری قلبی | سابقه سیگار | BMI | سطح HbA1c | سطح گلوکز خون | دیابت |
| 0 | 0.0002 | 44.23 | 0 | 0 | 3.9 | 28.23 | 5.51 | 137.38 | 0.07 |
| 1 | 1 | 38.48 | 0 | 0 | 1.93 | 26.71 | 5.51 | 137.25 | 0.07 |
| 2 | 0.4209 | 60.2 | 1 | 0 | 2.82 | 30.89 | 5.81 | 149.48 | 0.26 |
| 3 | 0 | 35.8 | 0 | 0 | 0.26 | 25.84 | 5.44 | 134.67 | 0.03 |
| 4 | 0.6039 | 67.33 | 0.23 | 1 | 2.44 | 29.3 | 5.88 | 152.14 | 0.32 |

**نتایج و تحلیل خوشه‌بندی سلسله مراتبی:**

* دندروگرام خوشه‌بندی سلسله مراتبی رسم شد.
* با توجه به دندروگرام، به نظر میرسد تعداد خوشه 4 به عنوان تعداد مناسب خوشه‌ها انتخاب شد. البته از روی روش قبلی تعداد خوشه 5 و 9 را نیز انجام دادیم.

**تحلیل خوشه‌ها (سلسله مراتبی):**

|  |  |  |
| --- | --- | --- |
| تعداد خوشه‌ها (n) | امتیاز سیلوئت | شاخص کالینسکی-هاراباس |
| 4 | 0.1799 | 943.08 |
| 5 | 0.21 | 975.92 |
| 9 | 0.2244 | 848.13 |

همچنین برای روش اصلی این بخش که با ان مساوی 4 انجام شده بود:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| خوشه | جنسیت | سن | فشار خون | بیماری قلبی | سابقه سیگار | BMI | سطح HbA1c | سطح گلوکز خون | دیابت |
| 0 | 0.06 | 41.62 | 0.09 | 0 | 2.23 | 27.44 | 5.41 | 132.92 | 0 |
| 1 | 1 | 35.89 | 0 | 0 | 1.8 | 25.69 | 5.39 | 132.13 | 0 |
| 2 | 0.6 | 66.92 | 0.24 | 1 | 2.48 | 29.76 | 6.02 | 146.83 | 0.29 |
| 3 | 0.48 | 59.42 | 0.26 | 0 | 2.76 | 31.95 | 7.01 | 197.58 | 1 |

که البته در نهایت مورد با ان مساوی 5 را به عنوان نتیجه نهایی در نظر گرفتیم:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| خوشه | جنسیت | سن | فشار خون | بیماری قلبی | سابقه سیگار | BMI | سطح HbA1c | سطح گلوکز خون | دیابت |
| 0 | 0.02 | 40 | 0 | 0 | 2.19 | 27.21 | 5.4 | 132.89 | 0 |
| 1 | 1 | 35.89 | 0 | 0 | 1.8 | 25.69 | 5.39 | 132.13 | 0 |
| 2 | 0.6 | 66.92 | 0.24 | 1 | 2.48 | 29.76 | 6.02 | 146.83 | 0.29 |
| 3 | 0.48 | 59.42 | 0.26 | 0 | 2.76 | 31.95 | 7.01 | 197.58 | 1 |
| 4 | 0.46 | 59.37 | 1 | 0 | 2.58 | 29.92 | 5.47 | 133.2 | 0 |

**مقایسه دو روش خوشه‌بندی:**

**معیارهای ارزیابی:**

برای ارزیابی کیفیت خوشه‌بندی، از دو معیار امتیاز سیلوئت و شاخص کالینسکی-هاراباس استفاده شد. نتایج به شرح زیر است:

|  |  |  |
| --- | --- | --- |
| معیار | K-means (k=5) | سلسله مراتبی (k=5) |
| امتیاز سیلوئت | 0.197 | 0.21 |
| شاخص کالینسکی-هاراباس | 19326 | 976 |

همانطور که مشاهده می‌شود، خوشه‌بندی سلسله مراتبی شاخص کالینسکی-هاراباس پایینتری دارد که نشان‌دهنده‌ی جدایی بدتر خوشه‌ها است، در حالی که K-means امتیاز سیلوئت پایین تری دارد که نشان‌دهنده‌ی انسجام داخلی بدتر خوشه‌ها است.

با توجه به این معیارها، K-means برای این داده‌ها به طور واضح روش مناسب‌تری به نظر می‌رسد. دلیل اصلی، شاخص کالینسکی-هاراباس بسیار بالاتر آن است که نشان‌دهنده‌ی جدایی بسیار بهتر خوشه‌ها است. اگرچه امتیاز سیلوئت در خوشه‌بندی سلسله مراتبی کمی بالاتر است، اما این تفاوت در مقابل تفاوت چشمگیر در شاخص کالینسکی-هاراباس ناچیز است.

**طبقه‌بندی:**

در این بخش، از الگوریتم‌های طبقه‌بندی زیر برای پیش‌بینی متغیر هدف diabetes استفاده شده است:

* **درخت تصمیم: (Decision Tree)** این الگوریتم یک مدل درختی از تصمیمات ایجاد می‌کند که بر اساس ویژگی‌ها، نمونه‌ها را به کلاس‌های مختلف دسته‌بندی می‌کند.
* **ماشین بردار پشتیبان: (SVM)** این الگوریتم با پیدا کردن یک ابرصفحه (hyperplane) که بیشترین فاصله را با نزدیک‌ترین نقاط داده از هر کلاس داشته باشد، داده‌ها را دسته‌بندی می‌کند.
* **k-نزدیکترین همسایه: (KNN)** این الگوریتم بر اساس شباهت بین نمونه‌ها، آن‌ها را دسته‌بندی می‌کند. برای هر نمونه جدید، k نزدیکترین همسایه آن در داده‌های آموزشی پیدا می‌شود و کلاس نمونه جدید بر اساس کلاس همسایه‌ها تعیین می‌شود.
* **رگرسیون لجستیک: (Logistic Regression)** این الگوریتم یک مدل خطی برای طبقه‌بندی باینری است که احتمال تعلق یک نمونه به یک کلاس خاص را پیش‌بینی می‌کند.

**نتایج و تحلیل طبقه‌بندی:**

در این بخش، عملکرد مدل‌های مختلف دسته‌بندی شامل درخت تصمیم (Decision Tree)، ماشین بردار پشتیبان (SVM)، نزدیک‌ترین همسایه (KNN) و رگرسیون لجستیک (Logistic Regression) با یکدیگر مقایسه می‌شوند. هدف از این مقایسه، انتخاب بهترین مدل برای پیش‌بینی ریسک دیابت بر اساس داده‌های موجود است.

**معیارهای ارزیابی:**

برای ارزیابی عملکرد مدل‌ها از معیارهای زیر استفاده شد:

* **دقت (Accuracy):** نسبت پیش‌بینی‌های درست به کل پیش‌بینی‌ها.
* **بازخوانی (Recall):** نسبت نمونه‌های مثبت واقعی که به درستی توسط مدل شناسایی شده‌اند.
* **دقت (Precision):** نسبت نمونه‌هایی که به عنوان مثبت پیش‌بینی شده‌اند و واقعاً مثبت هستند.
* **F1-score:** میانگین هارمونیک بین دقت و بازخوانی که تعادل بین این دو معیار را نشان می‌دهد.

میانگین این معیارها در طی اعتبارسنجی متقابل (Cross-Validation) به شرح زیر است:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| مدل | میانگین دقت تست | میانگین بازخوانی تست | میانگین دقت تست | میانگین F1-score تست |
| درخت تصمیم | 0.9557 | 0.7351 | 0.7418 | 0.7383 |
| SVM | 0.8939 | 0.9128 | 0.4404 | 0.5941 |
| KNN | 0.9608 | 0.6327 | 0.8713 | 0.7331 |
| رگرسیون‌جستیک | 0.8863 | 0.8804 | 0.4198 | 0.5685 |

**مقایسه عملکرد مدل‌ها:**

با بررسی جدول بالا می‌توان نکات زیر را مشاهده کرد:

* **KNN بالاترین دقت را دارد،** و پس از آن درخت تصمیم (۰.۹۵۵۷) قرار دارد. این نشان می‌دهد که این دو مدل در پیش‌بینی کلی نمونه‌ها عملکرد خوبی دارند.
* **SVM و رگرسیون لجستیک دقت پایین‌تری دارند،** که نشان می‌دهد این دو مدل در مقایسه با دو مدل دیگر، خطاهای بیشتری در پیش‌بینی‌ها دارند.
* **از نظر بازخوانی، SVM بهترین عملکرد را دارد** و رگرسیون لجستیک نیز بازخوانی نسبتاً خوبی (۰.۸۸۰۴) دارد. این نشان می‌دهد که این دو مدل در شناسایی نمونه‌های مثبت واقعی (افراد مبتلا به دیابت) عملکرد بهتری دارند. اما دقت پایین آن‌ها نشان می‌دهد که تعداد زیادی از نمونه‌ها به اشتباه به عنوان مثبت پیش‌بینی می‌شوند (مثبت کاذب).
* **درخت تصمیم و KNN تعادل بهتری بین دقت و بازخوانی ارائه می‌دهند،** و F1-score بالاتری دارند. F1-score برای درخت تصمیم ۰.۷۳۸۳ و برای KNN ۰.۷۳۳۱ است.

**انتخاب مدل مناسب:**

با توجه به معیارهای ارزیابی، **مدل درخت تصمیم و KNN بهترین عملکرد را در این داده‌ها دارند.** KNN بالاترین دقت را دارد، اما درخت تصمیم تعادل بهتری بین دقت و بازخوانی ارائه می‌دهد و F1-score بالاتری دارد. اگر هدف اصلی ما پیش‌بینی دقیق کلیه نمونه‌ها باشد، KNN گزینه مناسب‌تری است. اما اگر شناسایی دقیق افراد مبتلا به دیابت (حساسیت) اهمیت بیشتری داشته باشد و تعادل بین دقت و بازخوانی نیز مهم باشد، درخت تصمیم گزینه بهتری است.

SVM و رگرسیون لجستیک به دلیل دقت پایین و عدم تعادل بین دقت و بازخوانی، برای این داده‌ها مناسب نیستند. دقت پایین این مدل‌ها نشان می‌دهد که تعداد زیادی از پیش‌بینی‌های آن‌ها نادرست است، و این موضوع در کاربردهای پزشکی می‌تواند بسیار مشکل‌ساز باشد.

**نتیجه‌گیری:**

در این بخش، عملکرد چهار مدل دسته‌بندی مختلف مقایسه شد. با توجه به معیارهای ارزیابی، مدل درخت تصمیم و KNN به عنوان مدل‌های مناسب برای پیش‌بینی ریسک دیابت شناسایی شدند. انتخاب نهایی بین این دو مدل بستگی به اولویت‌های خاص پروژه دارد. اگر دقت کلی مهم‌تر باشد، KNN انتخاب بهتری است و اگر تعادل بین دقت و بازخوانی و شناسایی دقیق افراد مبتلا به دیابت مهم‌تر باشد، درخت تصمیم گزینه مناسب‌تری خواهد بود.

# 

# بخش چهارم نتیجه گیری

**۵. نتیجه‌گیری:**

این گزارش به بررسی و مقایسه‌ی روش‌های مختلف یادگیری ماشین برای پیش‌بینی ریسک ابتلا به دیابت پرداخت. هدف از این مطالعه، انتخاب بهترین مدل برای دسته‌بندی افراد در معرض خطر و همچنین شناسایی گروه‌های مختلف افراد بر اساس ویژگی‌های آن‌ها با استفاده از خوشه‌بندی بود.

**نتایج کلیدی:**

* **دسته‌بندی:**
  + در بین مدل‌های دسته‌بندی، مدل‌های درخت تصمیم و KNN عملکرد بهتری داشتند.
  + KNN بالاترین دقت (۰.۹۶۰۸) را در پیش‌بینی کلی نمونه‌ها نشان داد، که نشان می‌دهد این مدل در اکثر موارد به درستی تشخیص می‌دهد که آیا فرد به دیابت مبتلا است یا خیر. با این حال، باید توجه داشت که دقت به تنهایی در داده‌های نامتوازن معیار کاملی نیست.
  + درخت تصمیم تعادل بهتری بین دقت (۰.۷۴۱۸) و بازخوانی (۰.۷۳۵۱) ارائه داد و F1-score آن (۰.۷۳۸۳) کمی بالاتر از KNN بود. این نشان می‌دهد که درخت تصمیم در شناسایی افراد مبتلا به دیابت (حساسیت) و همچنین درستی پیش‌بینی‌های مثبت (دقت) عملکرد قابل قبولی دارد.
  + SVM و رگرسیون لجستیک به دلیل دقت پایین و عدم تعادل بین دقت و بازخوانی، عملکرد ضعیف‌تری داشتند. بازخوانی بالای SVM و رگرسیون لجستیک و دقت پایینشان نشان از تعداد زیاد مثبت کاذب در پیش‌بینی‌هایشان بود. این بدان معناست که این مدل‌ها تعداد زیادی از افراد سالم را به اشتباه به عنوان مبتلا به دیابت تشخیص می‌دهند که در کاربردهای پزشکی می‌تواند پیامدهای جدی داشته باشد.
* **خوشه‌بندی:**
  + در مقایسه بین K-means و خوشه‌بندی سلسله مراتبی با k=5، K-means به دلیل شاخص کالینسکی-هاراباس بسیار بالاتر (۱۹۳۲۶ در مقابل ۹۷۶)، جدایی خوشه‌های بهتری را نشان داد و به عنوان روش مناسب‌تر انتخاب شد. این نشان می‌دهد که خوشه‌های ایجاد شده توسط K-means فشرده‌تر و از یکدیگر دورتر هستند.
  + امتیاز سیلوئت پایین در هر دو روش خوشه‌بندی نشان می‌دهد که ساختار خوشه‌ای داده‌ها ممکن است پیچیده باشد و نیاز به بررسی بیشتر دارد. این موضوع می‌تواند ناشی از همپوشانی بین خوشه‌ها یا وجود خوشه‌هایی با شکل‌های غیر کروی باشد.

**پیشنهادات برای بهبود عملکرد مدل‌ها:**

با توجه به نتایج به دست آمده و تحلیل‌های انجام شده، پیشنهادات زیر برای بهبود عملکرد مدل‌ها ارائه می‌شود:

* **برای مدل‌های دسته‌بندی:**
  + **مدیریت داده‌های نامتوازن:** با توجه به احتمال وجود داده‌های نامتوازن (تعداد نمونه‌های افراد مبتلا به دیابت احتمالاً کمتر از افراد سالم است)، استفاده از روش‌هایی مانند SMOTE برای بیش‌نمونه‌برداری از کلاس اقلیت یا کم‌نمونه‌برداری از کلاس اکثریت توصیه می‌شود. این کار می‌تواند به بهبود بازخوانی و F1-score مدل‌ها، به خصوص درخت تصمیم و KNN، کمک کند.
  + **مهندسی ویژگی:** بررسی و انتخاب ویژگی‌های مرتبط‌تر و ایجاد ویژگی‌های جدید با استفاده از دانش دامنه می‌تواند به بهبود عملکرد مدل‌ها کمک کند. به عنوان مثال، ترکیب ویژگی‌ها یا تبدیل آن‌ها به فرمت‌های دیگر می‌تواند مفید باشد. همچنین، بررسی ارتباط بین ویژگی‌ها و هدف (دیابت) می‌تواند به انتخاب ویژگی‌های مؤثرتر کمک کند.
  + **تنظیم ابرپارامترها:** استفاده از روش‌هایی مانند Grid Search یا Random Search برای یافتن بهترین مقادیر ابرپارامترها برای هر مدل، به خصوص برای SVM که عملکرد نسبتا ضعیفی داشت، می‌تواند تأثیر زیادی داشته باشد.
  + **بررسی مدل‌های دیگر:** مدل‌های پیشرفته‌تر مانند Random Forest، Gradient Boosting (مانند XGBoost یا LightGBM) و شبکه‌های عصبی می‌توانند برای مقایسه و احتمالاً بهبود عملکرد مدل‌ها مورد بررسی قرار گیرند. این مدل‌ها معمولاً در مسائل پیچیده‌تر و داده‌های بزرگ‌تر عملکرد بهتری دارند.
* **برای خوشه‌بندی:**
  + **بررسی تعداد خوشه‌های مختلف (k):** با توجه به امتیاز سیلوئت پایین، بررسی مقادیر مختلف k و استفاده از روش‌هایی مانند روش Elbow Method یا Silhouette Analysis برای یافتن تعداد بهینه‌ی خوشه‌ها توصیه می‌شود. شاید k=5 بهترین انتخاب نباشد و تعداد خوشه‌های بیشتری یا کمتری برای داده‌های شما مناسب‌تر باشد.
  + **استفاده از الگوریتم‌های دیگر خوشه‌بندی:** الگوریتم‌هایی مانند DBSCAN که به شکل خوشه‌ها حساس نیستند و می‌توانند خوشه‌های با شکل‌های مختلف را شناسایی کنند، می‌توانند مفید باشند. این الگوریتم‌ها به خصوص در مواردی که داده‌ها ساختار خوشه‌ای غیر کروی دارند، عملکرد بهتری نشان می‌دهند.
  + **پیش‌پردازش داده‌ها:** نرمال‌سازی یا استانداردسازی داده‌ها برای جلوگیری از تأثیر مقیاس متفاوت ویژگی‌ها بر عملکرد خوشه‌بندی ضروری است. حذف داده‌های پرت نیز می‌تواند به بهبود نتایج کمک کند.
  + **کاهش ابعاد:** استفاده از روش‌های کاهش ابعاد مانند PCA می‌تواند به کاهش پیچیدگی داده‌ها و بهبود عملکرد خوشه‌بندی کمک کند. این روش‌ها می‌توانند ویژگی‌های اصلی داده‌ها را در فضای با ابعاد کمتر حفظ کنند و در عین حال نویز و اطلاعات اضافی را کاهش دهند.

**نتیجه‌گیری نهایی:**

در این مطالعه، روش‌های مختلف یادگیری ماشین برای پیش‌بینی ریسک دیابت و شناسایی گروه‌های مختلف افراد مورد بررسی قرار گرفتند. در بخش دسته‌بندی، مدل‌های درخت تصمیم و KNN به عنوان گزینه‌های مناسب‌تر شناسایی شدند، اما نیاز به بهبود عملکرد آن‌ها با استفاده از روش‌های پیشنهادی وجود دارد. در بخش خوشه‌بندی، K-means به دلیل جدایی بهتر خوشه‌ها، روش مناسب‌تری تشخیص داده شد، اما بررسی بیشتر ساختار خوشه‌ای داده‌ها و استفاده از روش‌های دیگر خوشه‌بندی توصیه می‌شود. با اعمال پیشنهادات ارائه شده، می‌توان انتظار داشت که دقت پیش‌بینی‌ها و کیفیت خوشه‌بندی بهبود یابد و تحلیل دقیق‌تری از داده‌ها ارائه شود.