

**گزارش تکلیف اول درس یادگیری ماشین کاربردی**

**استاد درس: احمدی**

تهیه کنندگان:

سید نیما محمودیان

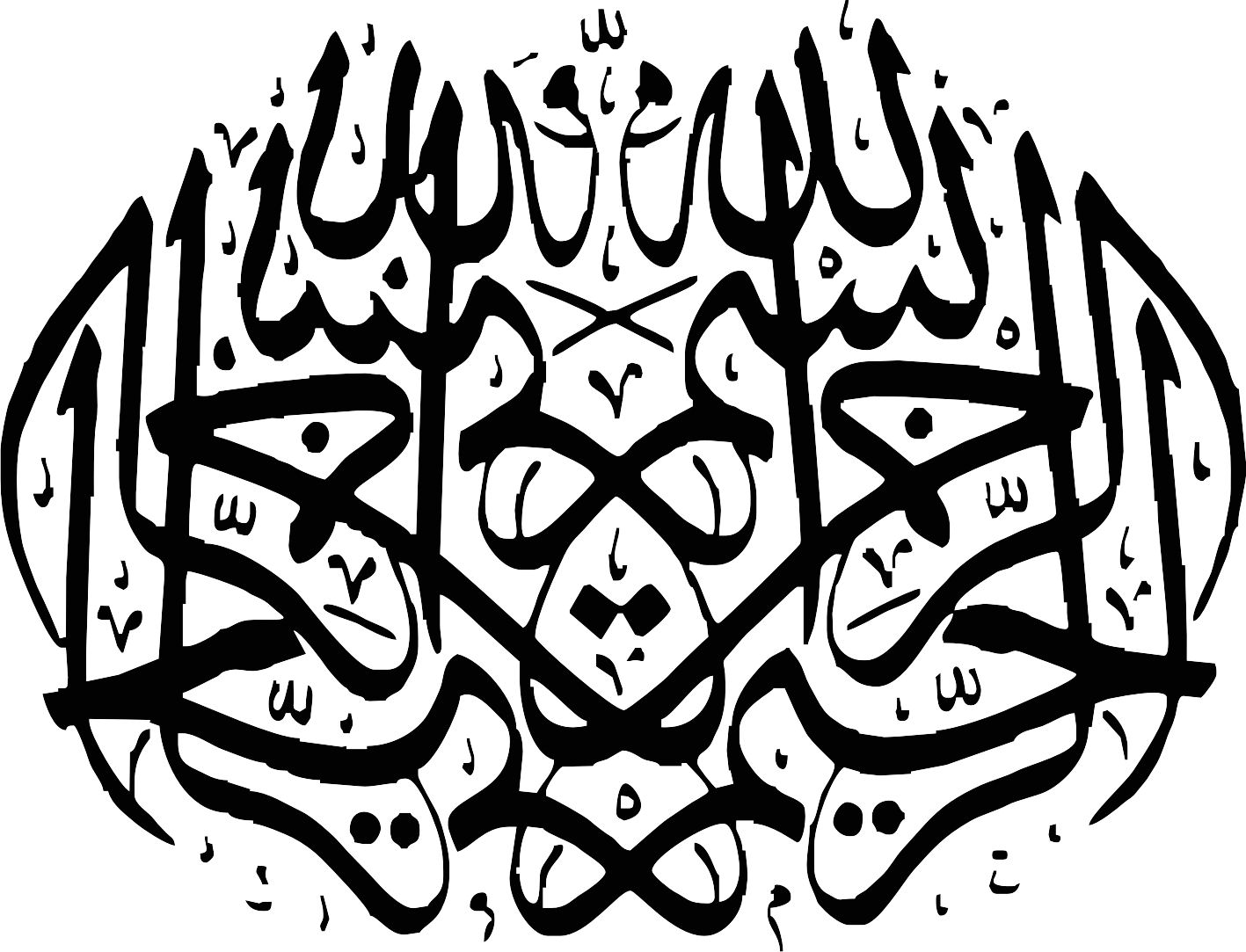
402125005

معصومه بهبهانی‌فر

402191003

محسن فراغه

401225002



فهرست مطالب

[1-پاسخ سوال اول 1](#_Toc163208856)

[1-1-لود کردن دیتاست، نمایش ده ردیف به صورت تصادفی، و بررسی کلی داده‌ها 1](#_Toc163208857)

[2-1-مدیریت مقادیر گمشده 5](#_Toc163208858)

[3-1-مصور سازی داده‌ها 8](#_Toc163208859)

[4-1-مدیریت داده‌های پرت 14](#_Toc163208860)

[5-1-مهندسی ویژگی‌ها 16](#_Toc163208861)

[2-پیش‌پردازش تصویر 18](#_Toc163208862)

[1-2- gray scale کردن عکس‌ها 20](#_Toc163208863)

[2-2-تنظیم روشنایی و کنتراست تصاویر 21](#_Toc163208864)

[3-2-نرمال سازی تصاویر 23](#_Toc163208865)

[3-پیش‌پردازش متن 25](#_Toc163208866)

[1-3-خواندن داده‌ها از پیکره همشهری 25](#_Toc163208867)

[2-3-پیش‌پردازش متن‌ها 29](#_Toc163208868)

[3-3-TF-IDF 32](#_Toc163208869)

[4-3-مصورسازی داده‌ها 33](#_Toc163208870)

[4-پیوست 40](#_Toc163208871)

فهرست شکل‌ها

**No table of figures entries found.**

# 1-چکیده

در این مطالعه تحلیل بر روی مجموعه داده[[1]](#footnote-1) دیابت انجام شده است. داده­های مورد بررسی از [[2]](#footnote-2)UCI گرفته شده است. پس از معرفی مجموعه داده، فرآیند آماده سازی داده­ها و بررسی مقادیر گمشده و پرت را در دستور کار داریم، سپس به استاندارد سازی داده­ها می­پردازیم. در ادامه کار با استفاده از روش جست و جوی رو به جلو ویژگی­های کاربردی تر را از بین ویژگی­های موجود انتخاب می­کنیم. سپس معیارهای دقت، حساسیت و اختصاصیت را برای این انتخاب ویژگی­ها به دست می­آوریم. در ادامه با استفاده از روش KNN [[3]](#footnote-3) دسته بندی داده­ها را انجام می­دهیم. معیارهای نام برده در بالا را دوباره برای مدل ارائه شده محاسبه می­کنیم و تمامی مراحل را با استفاده از روش دسته­بندی بیزی در دستور کار خواهیم داشت.

مسئله دوم مورد بررسی در این پژوهش، شامل بررسی و به دست آوردن مرز تصمیم بین دو کلاس با استفاده از روش­های نزدیکترین همسایگی و پرسپترون می­باشد.

# 2**-معرفی مجموعه داده**

مجموعه داده مورد بررسی، اطلاعات 768 فرد را نمایش می­دهد، هدف مجموعه داده این است که بر اساس اندازه‌گیری‌های تشخیصی خاص موجود در مجموعه داده، پیش‌بینی تشخیصی اینکه آیا بیمار مبتلا به دیابت است یا خیر را انجام دهد. در جمع­آوری این مجموعه داده محدودیت­هایی نیز ایجاد شده است. این مجموعه داده اطلاعات زنان بالای 21 سال از میراث هندی پیما هستند. توضیحات هر یک از ویژگی مورد بررسی عبارتند از:

1. بارداری: تعداد دفعاتی که فرد باردار شده است.
2. گلوکز: غلظت کلوگز پلاسما در یک آزمایش دو ساعته
3. فشار خون: فشار خون دیاستولیک
4. ضخامت پوست: ضخامت چین­های پوستس سه سر بازو بر حسب میلی متر
5. انسولین: سرم دو ساعته انسولین
6. شاخص توده بدنی BMI[[4]](#footnote-4): وزن بر حسب کیلوگرم / قد بر حسب متر به توان دو
7. DiabetesPedigreeFunction: تابعی از سابقه خانوادگی فرد از ابتلا به دیابت
8. سن: سن افراد شرکت کننده

# 3- فراخوانی داده­ها

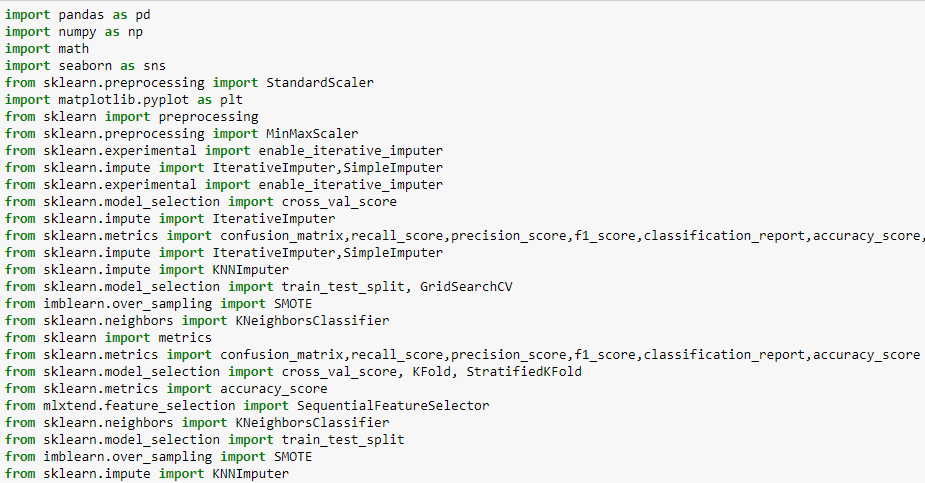
## 1-3 سوال

**مساله یک**. مجموعه داده pima-indian-diabetes از مخزن دادههای استاندارد دانشگاه کالیفرنیا ( UCI Machine Learning Repository)مورد نظر است. سوالات زیر را پاسخ دهید.

الف) بعد از پیش پردازش داده­ها، با استفاده از روش جست و جوی رو به جلو، بهترین ترکیب ویژگی­ها را به دست آورید. از روش نزدیکترین همسایگی جهت دسته بندی استفاده کنید. بدین منظور از روش Hold out استفاده کنید. مقدارهای دقت، حساسیت و اختصاصیت را گزارش کنید.

## 2-3 کدهای اولیه و فراخوانی داده­ها

در ابتدا کتابخانه­های مورد نیاز را فراخوانی می­کنیم. در زیر بخشی از آن­ها را مشاهده می­کنید:



شکل 2-1 کد فراخوانی کتابخانه­ها

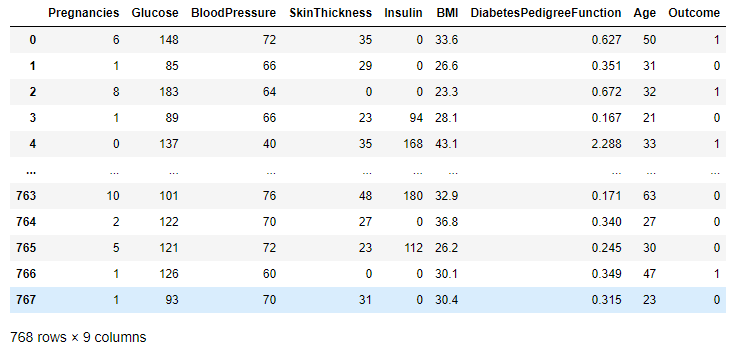
با توجه به این مهم، که برای بررسی این مجموعه داده فایل CSV[[5]](#footnote-5) مربوط را دانلود کرده­ایم، برای فراخوانی این مجموعه داده خواهیم داشت:



شکل 2-2 کد فراخوانی مجموعه داده

در ادامه یک نمای کلی از مجموعه داده را نمایش می­دهیم:

جدول 2-1 نمای کلی مجموعه داده



## 3-3 تحلیل توصیفي داده­ها

در گذشته همانگونه که در شکل 2-2 مشاهده کردید، یک قالب داده تحت عنوان “df” از کل مجموعه داده­ای که در اختیار داشتیم، ایجاد کردیم. برای تسلط بیشتر بر روی مجموعه داده اقدامات زیر را انجام می­دهیم:

df.head : برای مشاهده پنج سطر اول از مجموعه داده استفاده می­شود.

df.tail: برای مشاهده پنج سطر آخر از مجموعه داده استفاده می­شود.

df.nunique: تعداد مقادیر منحصر به فرد را در هر ستون از چارچوب داده برمیگرداند. این به ما نشان می­دهد که در هر ستون چند مقدار مختلف وجود دارد و داده­ها چقدر متنوع هستند.

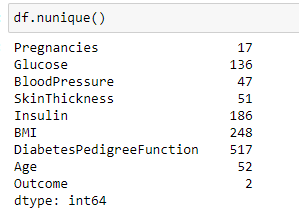
df.describe: تعداد داده­های هر ستون، مقادیر میانگین، انحراف معیار، کمترین و بیشترین مقدار هر ستون و چارک­ها را نمایش می­دهد.

df.info : خلاصه­ای از داده­ها شامل فهرست، نام ستون­ها، انواع داده­ها، مقادیر غیرتهی، میزان استفاده از حافظه و سایر اطلاعات را چاپ می­کند. این یک نمای کلی از چارچوب داده و ویژگی­های آن به ما می­دهد.

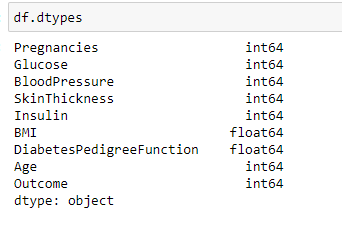
df.shape: تعداد ردیف­ها و ستون­ها را به ما برمی­گرداند.

df.dtype: نوع هر داده را مشخص می­کند.

برخی از خروجی­های کدهای بالا را مشاهده می­کنید:

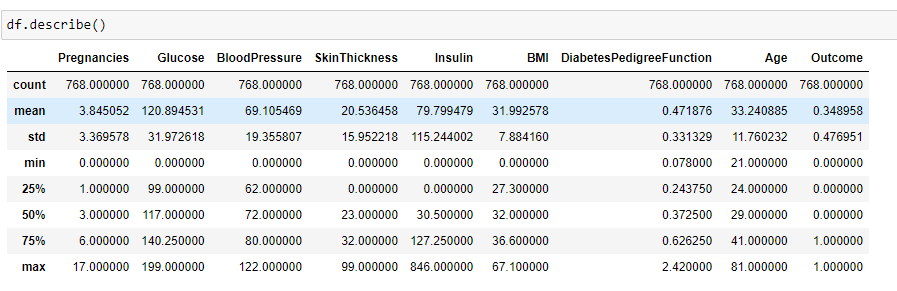


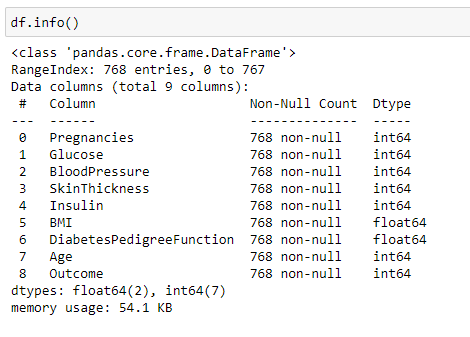
شکل 2-3 خروجی کد df.nunique



شکل 2-4 خروجی کد df.dtype

جدول 2-2 خروجی کد df.describe



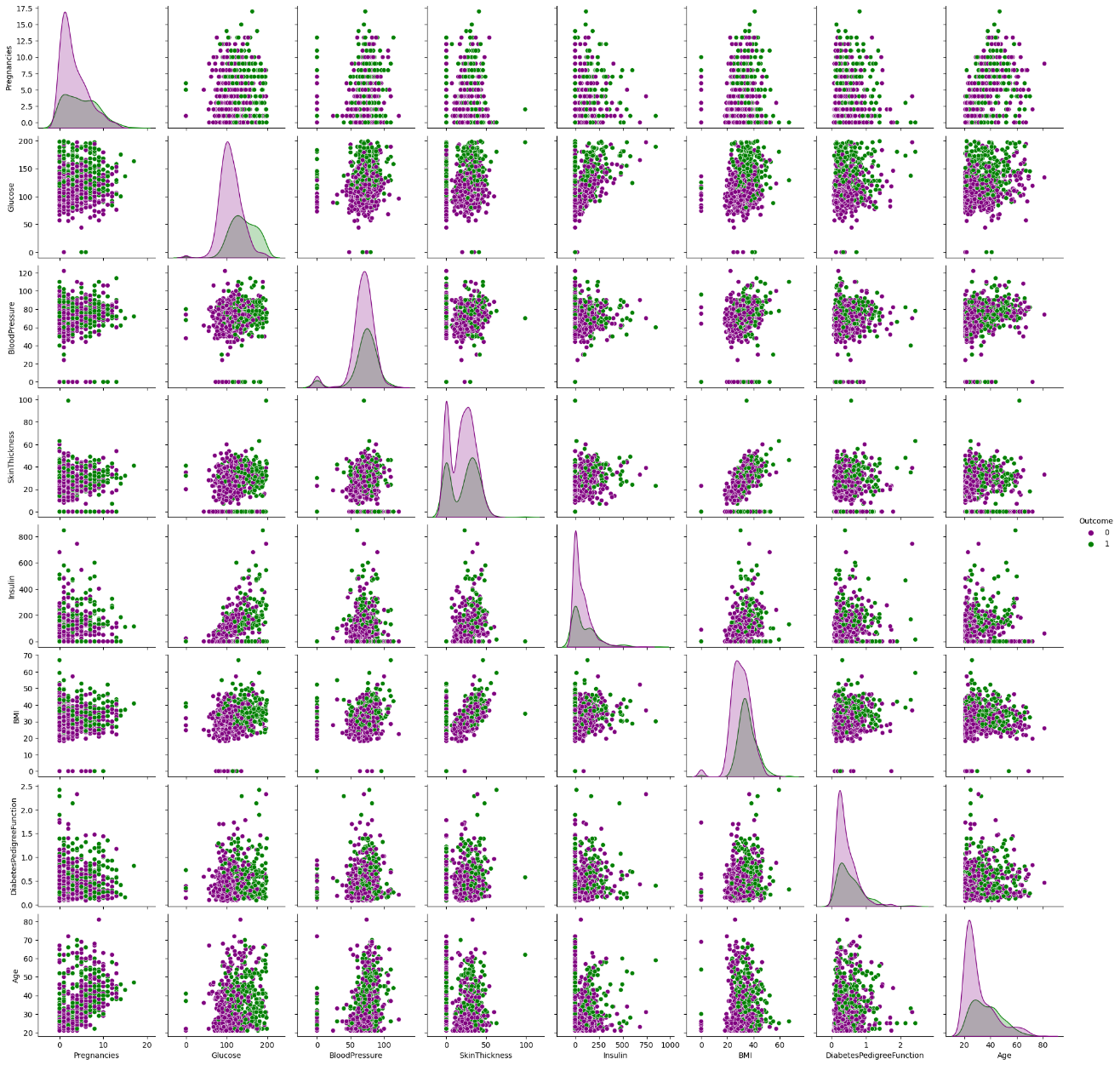


شکل 2-5خروجی کد df.info

در ادامه برای شناخت بهتر مجموعه داده و ارتباط دو به دو هر یک از ویژگی­ها با یکدیگر نمودارهای زیر را رسم کرده­ایم:

### 1-3-3 نمودار Pairplot

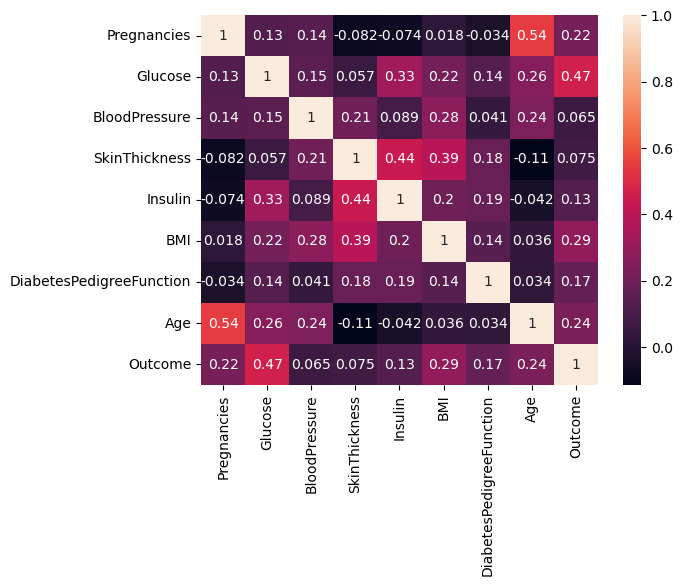
این نمودار که به عنوان ماتریس پراکندگی نیز شناخته می شود، ماتریسی از نمودارها می­باشد که امکان تجسم رابطه بین هر جفت متغیر در یک مجموعه داده را فراهم می­کند. هر دو نمودار هیستوگرام و پراکندگی را ترکیب می کند و یک نمای کلی منحصر به فرد از توزیع­ها و همبستگی­های مجموعه داده ارائه می­دهد.



شکل 2-6 Pairplot

### 2-3-3 نمودار Heatmap

همانگونه که در شکل پیداست نقشه­های حرارتی برای نشان دادن روابط بین دو متغیر استفاده می­شود که یکی در هر محور ترسیم شده است. با مشاهده چگونگی تغییر رنگ سلول­ها در هر محور، می­توانید مشاهده کنید که آیا الگوهایی در ارزش یک یا هر دو متغیر وجود دارد یا خیر.

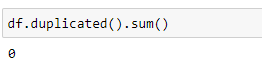


شکل 2-7 نمودار Heatmap

## 4-3 پیش پردازش داده­ها

### 1-4-2پیدا کردن داده­های تکراری

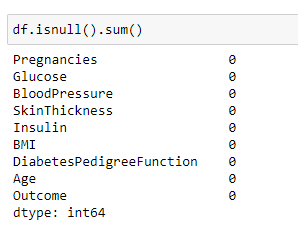
برای بررسی مجموعه داده باید اطمینان حاصل کنیم که هیچ داده تکراری در مجموعه داده ما وجود ندارد. اگر ستون­های کاملا مشابهی در مجموعه داده ما وجود دارد، باید حذف شوند. که در مجموعه داده مورد بررسی ما طبق کد دستوری زیر، هیچ دو ستون مشابهی موجود نمی­باشد:



شکل 2-8 مشخص کردن داده­های تکراری

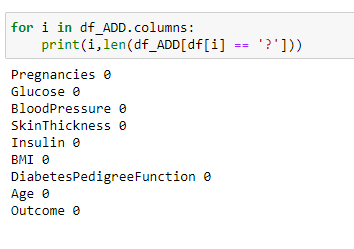
### 2-4-3 پیدا کردن مقادیر گمشده

یکی از مهم­ترین چالش­هایی که در آماده سازی داده­ها با آن سر و کار داریم، داشتن مقادیر گمشده می­باشد. مقادیر گمشده ممکن است در پیش­بینی­های آینده یا در روند کار برای ما مشکلاتی را ایجاد کنند، از این رو شناسایی آن­ها حائز اهمیت است. کد زیر تعداد داده­های گم شده را نمایش می­دهد:



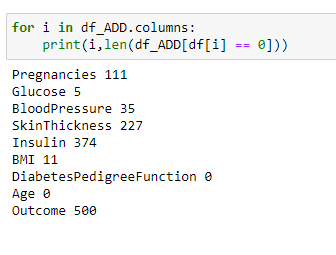
شکل 2-9 نمایش مقادیر گمشده

همانگونه که در شکل 2-10 مشاهده می­شود، هیچگونه مقدار گمشده­ای در این مجموعه داده موجود نمی­باشد. اما نکته قابل تامل در این بخش شناسایی داده­هایی با مقدار " ؟ " و یا " 0 " می­باشد. همان­طور که متوجه شده­اید با کد موجود در شکل 2-10 تنها مقادیری که وجود ندارند، مشخص می­شود. این در حالی است که گاهی وجود " 0 " برای یک شاخص معنی ندارد و حضور " ؟ " نیز در مجموعه داده به معنای عدم وجود مقدار است.



شکل 2-10 نمایش تعداد مقادیر " ؟ "

همانگونه که مشاهده می­شود، هیچگونه داده­ای با مقدار " ؟ " وجود ندارد. حال تعداد " 0 " را نیز بررسی می­کنیم.



شکل 2-11 نمایش مقادیر " 0 "

همانگونه که مشاهده می­شود برای شاخص­های بارداری، گلوکز، فشار خون، ضخامت پوست، انسولین، BMI و خروجی مجموعه داده، مقادیر صفر داریم. در این بین برای شاخص­های گلوکز، فشار خون، ضخامت پوست، انسولین و BMI ، مقدار صفر، مقداری نامعتبر و نشدنی است، پس نتیجه می­گیریم که این مقادیر در اصل ناموجود یا گمشده بوده­اند که با جای خالی گذاشتن، مقدارصفر را جایگزین کرده­اند.

پس از شناسایی داده­های گمشده، شناسایی داده­های پرت (نویز[[6]](#footnote-6)) را در دستور کار خواهیم داشت.

### 3-4-3 پیدا کردن مقادیر پرت

نویز، داده­هایی را شامل می­شود که به هر دلیلی دارای مقادیر نادرست یا غلط باشند. مانند عکس­های تار در دوربین ثبت تخلف، تفاوت در واحد پول چند ویژگی مختلف

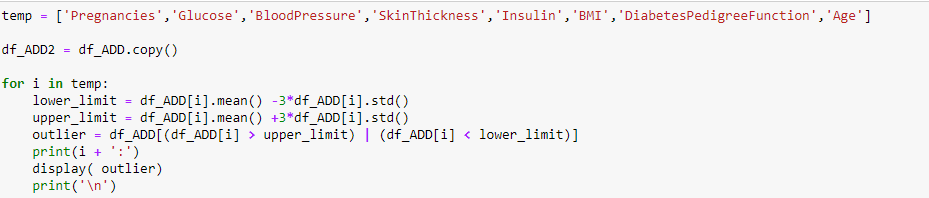
داده­های پرت نیز به داده­هایی گفته می­شود که در فا­صله­ی غیرعادی از سایر مقادیر داده در یک نمونه­ی تصادفی از یک

جمعیت مشاهده می­شود.

شناسایی نویزها و داده­های پرت امری حیاتی و تاثیرگذار در مبحث آماده سازی داده­ها می­باشد. از این رو باید در صدد رفع و شناسایی علت آن­ها باشیم. روش­های شناسایی این مشکلات عبارتند از:

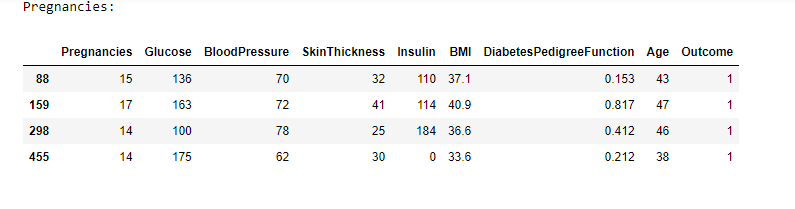
* شناسایی نقطه پرت بر اساس مفهوم آماری پراکندگی
* استفاده از روش­های خوشه بندی
* گسسته کردن داده­ها

در این پژوهش برای شناسایی مقادیر پرت از روش شش سیگما [[7]](#footnote-7) ، استفاده کرده*­­*ایم. این روش مقادیر مثبت بیشتر از 3 سیگما و مقادیر منفی کمتر از 3 سیگما را داده پرت شناسایی می­کند.

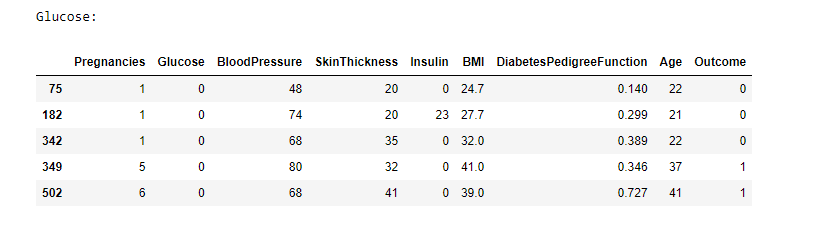


شکل2-12 کد مربوط به نمایش مقادیر پرت

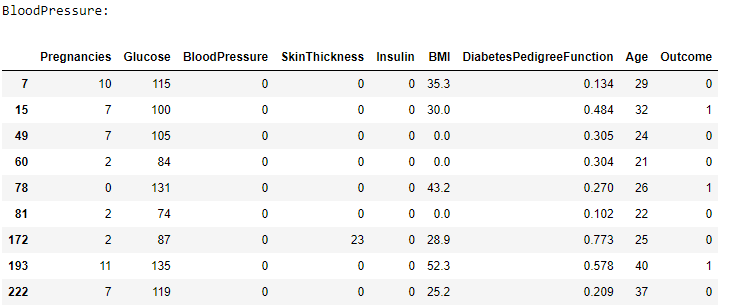
در زیر مقادیر پرت هر شاخص نمایش داده شده است:

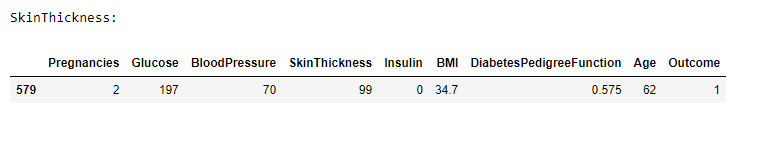
جدول 2-3 مقادیر پرت شاخص بارداری

جدول 2-4 مقادیر پرت شاخص گلوکز

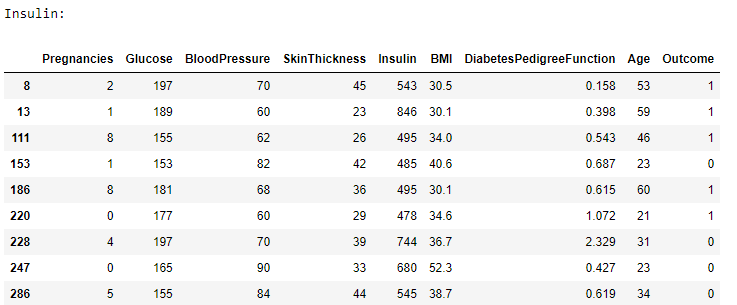


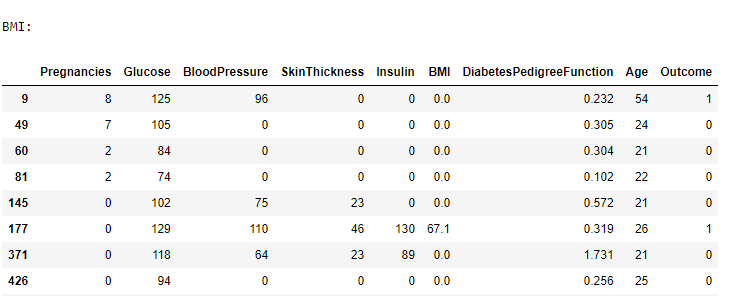
جدول 2-5 مقادیر پرت شاخص فشار خون

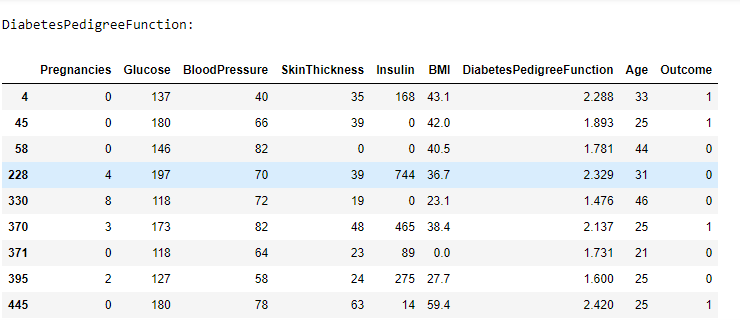


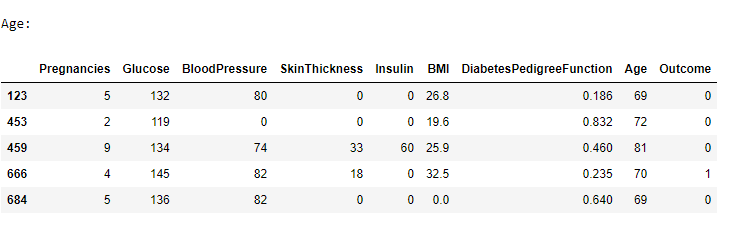
جدول 2-6 مقادیر پرت شاخص ضخامت پوست

جدول 2-7 بخشی از مقادیر پرت شاخص انسولین



جدول 2-8 مقادیر پرت شاخص BMI

جدول 2-9 مقادیر پرت شاخص DiabetesPedigreeFunction 

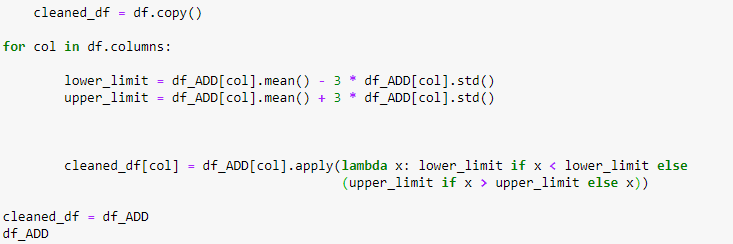
جدول 2-10 مقادیر پرت شاخص سن

### 4-3-3 رفع مشکل وجود داده­های پرت

برای رفع مشکل وجود داده­های پرت، چندین راه حل وجود دارد:

* حذف داده­های پرت
* جایگذاری داده­های پرت با مقدار میانگین
* جایگذاری داده­های پرت با مقادیر روی حد بالا و پایین

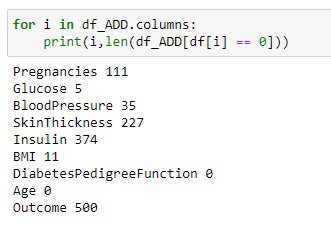
در این بخش داده­های بیشتر از حد را با مقادیر روی حد بالا و کمتر را با مقادیر روی حد پایین جایگذاری می­کنیم.



شکل 2-13 جایگذاری داده­ای پرت با مقادیر حد بالا و پایین

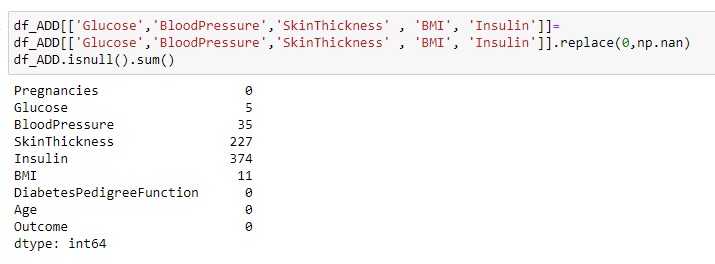
### 5-3-3 رفع مشکل وجود داده­های گمشده

در این بخش به رفع مشکل داده­های گمشده می­پردازیم. با توجه به جایگذاری داده­های پرت با مقادیر حد بالا و پایین، در ابتدا مجددا تعداد صفرهای موجود در هر شاخص را بررسی می­کنیم، زیرا ممکن است تعدادی از صفرها حذف و یا به تعداد آن­ها افزوده شده باشد.



شکل 2-14 نمایش مججد تعداد مقادیر " 0 "

با توجه به آنچه که از خروجی شکل 2-14 به دست آمده است، متوجه می­شویم که شاهد تغییر خاصی در مقادیر " 0 " نیستیم. حال تمامی مقادیر " 0 " را با مقدار nan جایگذاری می­کنیم.



شکل 2-15 جایگذاری مقادیر nan به جای " 0 "

از این بخش به بعد، باید به دنبال جایگذاری داده­های گمشده باشیم، روش­های مقابله با این مشکل عبارتند از :

* حذف مقادیر گمشده
* جایگذاری آن­ها با میانگین هر شاخص
* پر کردن داده­ها با مقادیر قبلی یا بعدی
* پر کردن با استفاده از مقادیر مشابه
* پر کردن با استفاده از مدل­های پیش­بینی

در این بخش، با استفاده از روش‌های Iterative ، Mean و Single به رفع این مشکل پرداخته ایم. روش IterativeImputer در ماشین لرنینگ یک روش پر کردن مقادیر گمشده در داده‌ها است.

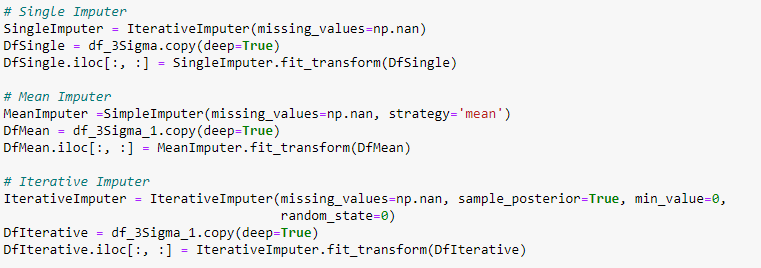
این روش، بر اساس روشی تکراری عمل می‌کند. ابتدا الگوریتم یک مدل پیش‌بینی را انتخاب می‌کند، مانند یک مدل رگرسیون یا یک مدل شبکه عصبی. سپس مقادیر گم شده را با استفاده از این مدل پیش‌بینی می‌کند.

مرحله بعدی این است که مقادیر پیش‌بینی شده را به عنوان مقادیر جدید برای مقادیر گم شده استفاده کنیم و این مقادیر را در داده‌ها جایگزین کنیم. با این کار، مقادیر گم شده جایگزین شده و یک نسخه جدید از داده‌ها به دست می‌آید. این فرآیند تکرار می‌شود و هر بار مدل پیش‌بینی بر اساس داده‌های جدید آموزش داده می‌شود. این روند تا زمانی ادامه پیدا می‌کند که مقادیر گم شده به اندازه کافی پر شوند یا معیاری مانند تعداد تکرارها مشخص شود.

روش IterativeImputer برای پر کردن مقادیر گمشده از اطلاعات موجود در داده‌ها بهره می‌برد. با استفاده از روش تکراری، مدل پیش‌بینی تلاش می‌کند تا از روابط و الگوهای موجود در داده‌ها استفاده کند و مقادیر مفقود را تخمین بزند.

Single Imputer روشی است که مقادیر گمشده را با استفاده از یک مقدار واحد برای هر ویژگی در مجموعه داده به حساب می‌آورد.

Mean Imputer محاسبه‌گر میانگین نوعی از محاسبه‌گر منفرد است که مقادیر گمشده را با استفاده از مقدار میانگین ویژگی محاسبه می‌کند.



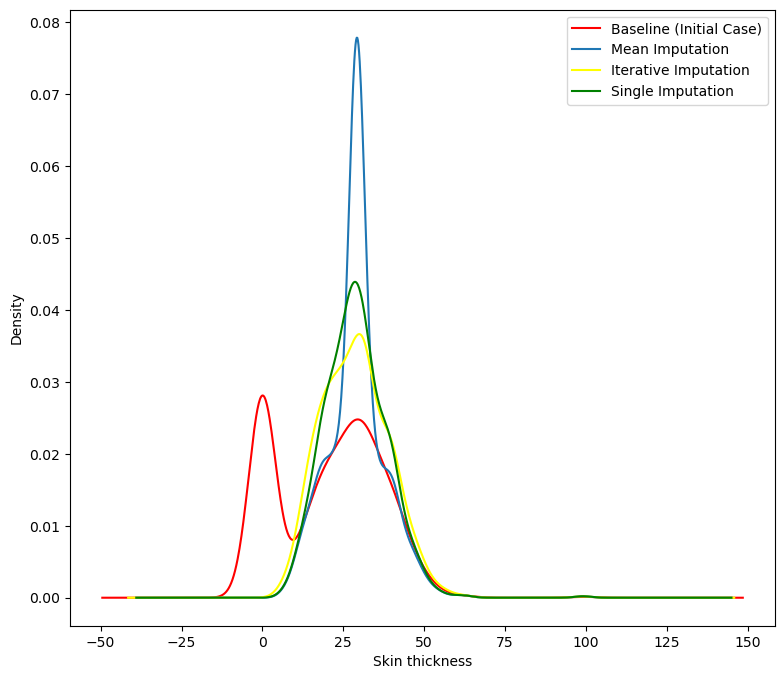
شکل 2-16 کد اجرا شده برای جایگذاری مقادیر گمشده



شکل 2-17 کد اجرا شده برای نمایش نمودار مقایسه­ای هر یک از روش­ها

#### 1-5-3-3 نمودار ویژگی ضخامت پوست

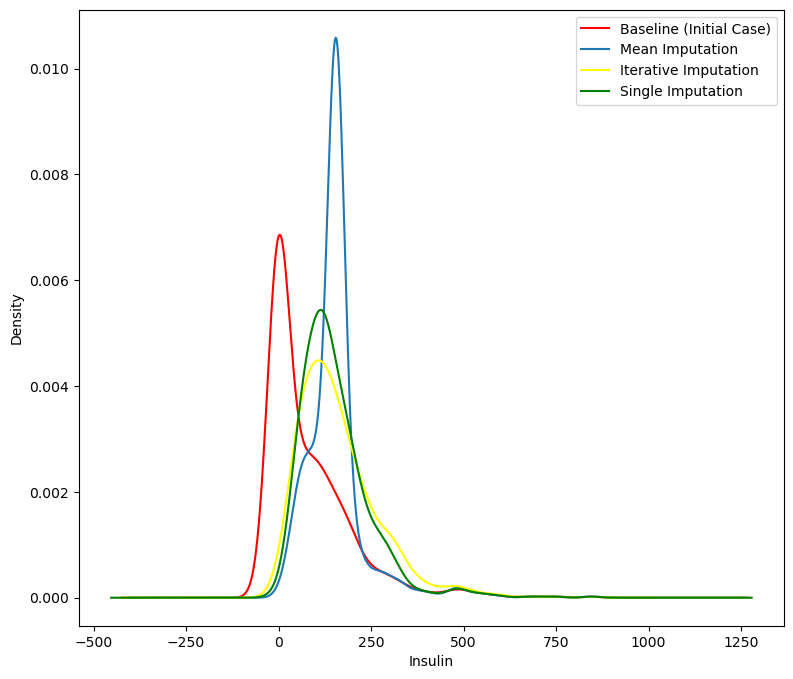
سه روش گفته شده برای پر کردن داده‌های گمشده را برای این ویژگی انجام دادیم، خط قرمز رنگ در نمودار خود داده اصلی هست. ما می‌خواهیم ببینیم کدوم یک از روش‌ها با این خط قرمز همپوشانی بیشتری دارد. در این بخش مشاهده می­شود که نمودار زرد رنگ تشابه بیشتری به نسبا بقیه نمودارها با نمودار اصلی دارد.



شکل 2-18 نمودار ضخامت پوست

#### 2-5-3-3 نمودار ویژگی انسولین

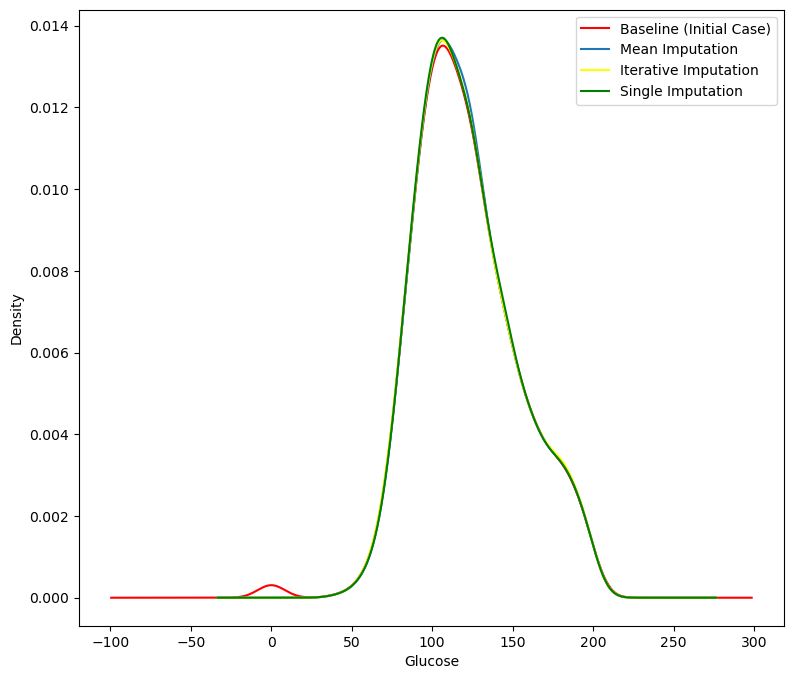
ویژگی انسولین را هم مانند قبل تحلیل می­کنیم. در این بخش نموداری که تشابه زیادی با نمودار قرمز رنگ داشته باشد موجود نیست.



شکل 2-19 نمودار انسولین

#### 3-5-3-3 نمودار ویژگی گلوکز

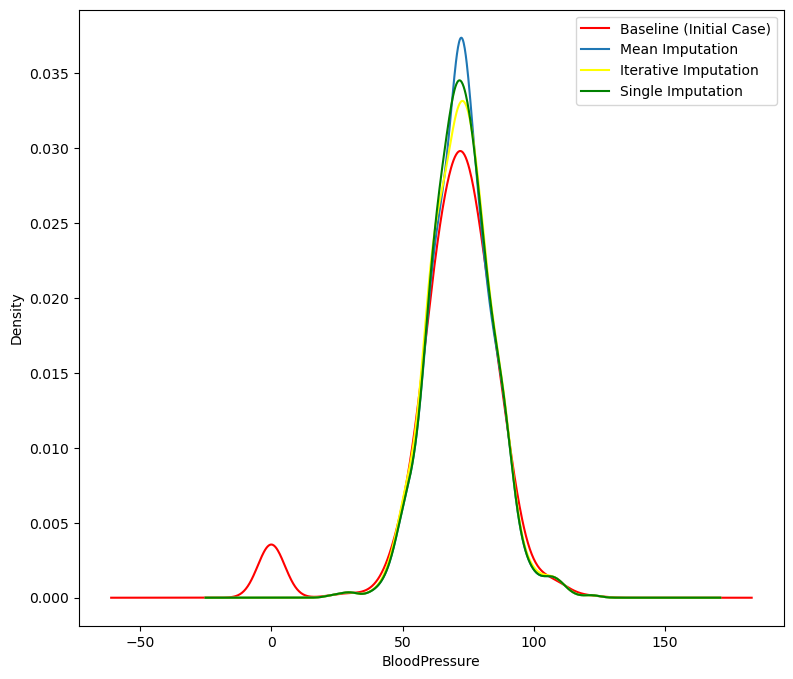
در این بخش مشاهده می­کنیم که تمامی روش­های جایگذاری بر روی نمودار اصلی تقریبا منطبق شده­اند. این مهم به علت تعداد کم داده­های گمشده برای این ویژگی می­باشد. اما با توجه و دقت بیشتر مشاهده می­کنیم که نمودار زرد رنگ نزدیکی بیشتری به نمودار اصلی دارد.



شکل 2-20 نمودار گلوکز

#### 4-5-3-3 نمودار ویژگی فشار خون

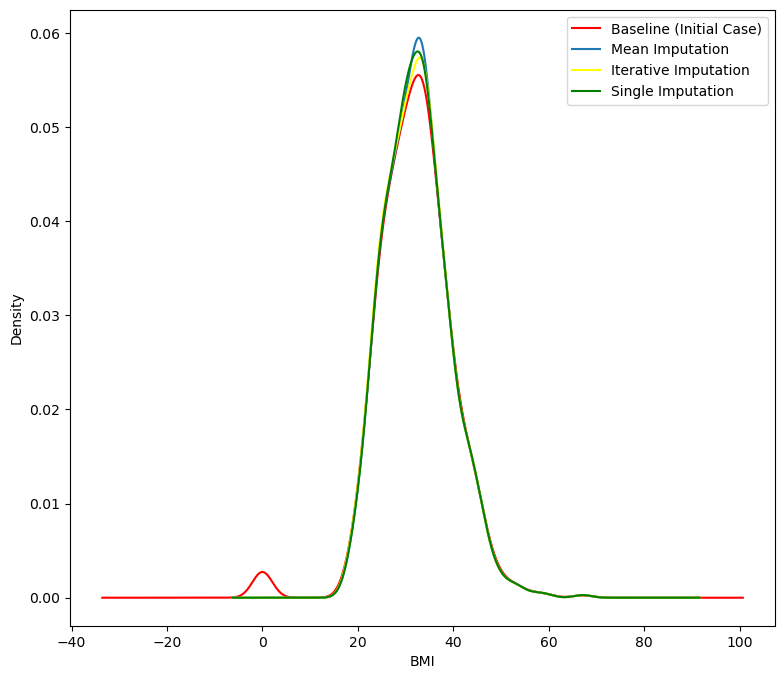
در این بخش نیز، تحلیل مانند قبل صورت می­گیرد. نزدیکترین نمودار به نمودار اصلی، نمودار زرد رنگ می­باشد.



شکل 2-21 نمودار فشارخون

### 5-5-3-2 نمودار ویژگی BMI

با مشاهده نمودار شکل 2-22 ، متوجه می­شویم که این ویژگی نیز عملکردی مشابه ویژگی­های قبل دارد. نزیکترین نمودار به نمودار اصلی، نمودار زرد رنگ می­باشد.

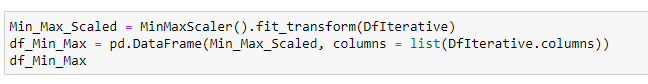


شکل 2-22 نمودار BMI

طبق مشاهدات صورت گرفته در این بخش، نزدیکترین مدل جایگذاری مقادیر گمشده به مجموعه داده اصلی روش IterativeImputer می­باشد.

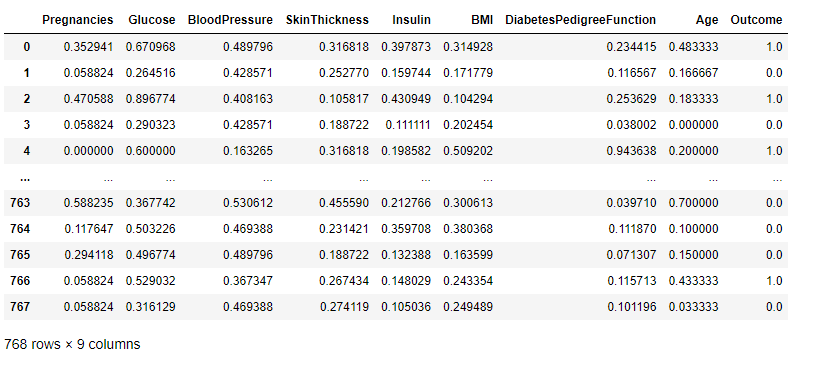
### 6-3-3 استاندارد سازی داد­ه­ها

در این بخش برای استانداردسازی داده ها از روش مین مکس استفاده کرده­ایم. این روش، نوعی الگوریتم بهینه سازی است که برای یافتن بهترین راه حل ممکن برای یک مسئله معین با جست و جو در تمام راه حل­های ممکن و انتخاب راه حل با بالترین ارزش یا کمترین هزینه استفاده می­شود. با به حداقل رساندن حداکثر هزینه یا به حداکثر رساندن حداقل مقدار یک مجموعه معین از پارامترها کار می­کند.



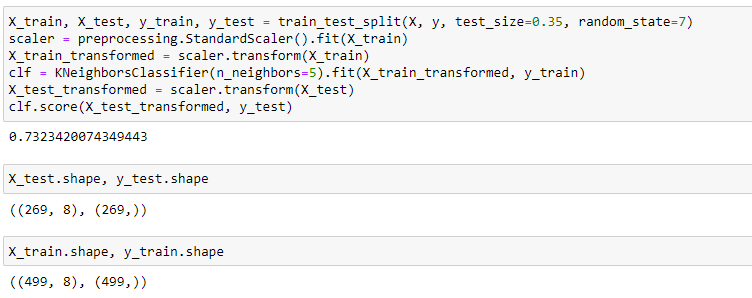
شکل 2-23 استاندارد سازی داده

جدول 2-11 داده­های استاندارد شده به روش مین مکس



### 7-3-3 **تقسیم بندی داده­ها به دو دسته آموزشی و آزمایشی**

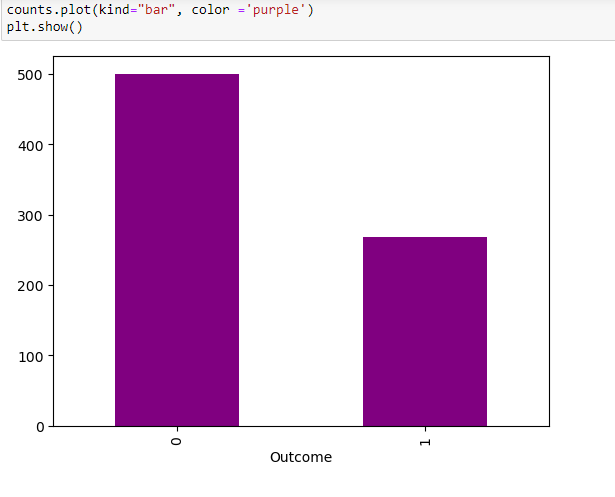
یکی از مهم­ترین کارها در پردازش داده، تقسیم بندی داده­ها به گروه­­های آموزشی و آزمایشی می­باشد. در این بخش 35 درصد از داده­ها را به دسته آزمایشی و 65 درصد را به آموزشی تقسیم بندی می­کنیم.

******

شکل 2-24 کد تقسیم داده­ها به دو دسته آموزشی و آزمایشی

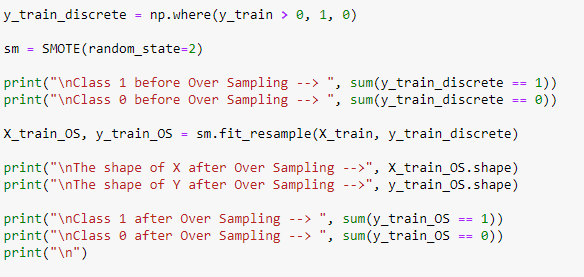
### 2-3-7-1 متعادل کردن داده­های آموزشی

شکل شماره 2-25 تعداد داده­های موجود در هر کلاس را نمایش می­دهد. با توجه به نمودار مربوط خواهیم دید که داده­ها متعادل نیستند.

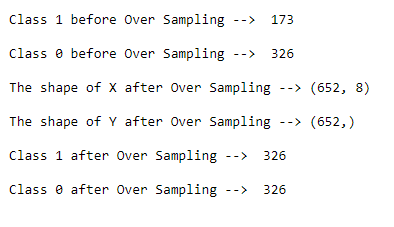
**

شکل 2-25 نمایش نمودار میله­ای داده­ها

همانگونه که می­دانیم برای متعادل سازی داده­ها، چندین روش وجود دارد. در این بخش ما از روش OverSampling برای متعادل سازی داده­ها استفاده کرده­ایم.



شکل 2-26 کد متعادل سازی مجموعه داده

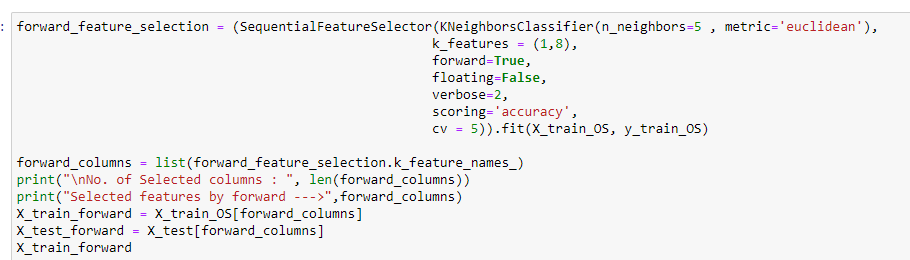


شکل 2-27 خروجی پس از متعادل سازی مجموعه داده

### 2-3-8 انتخاب ویژگی با روش جست و جو رو به جلو

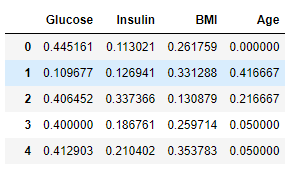
انتخاب ویژگی، به معنی انتخاب زیرمجموعه­ای از ویژگی­ها یا فاکتوهای مرتبط با مجموعه داده­ها می­باشد که بیشترین اطلاعات مفید را برای مدلسازی فراهم می­کنند. زمانی که از روش انتخاب ویژگی استفاده می­کنیم، تعداد ویژگی­های ورودی به مدل کاهش پیدا می­کند. دلیل استفاده از این روش این است که معمولاً مدل­های یادگیری ماشین برای مجموعه داده­ها با تعداد ویژگی­های بالا ، به دلیل ابعاد بالا و مواجه شدن با عملیات محاسباتی زیاد و در نهایت نتایج ضعیف، می­توانند با مشکل مواجه شوند. بنابراین، با کاهش تعداد ویژگی­ها، می­توانیم هر دو مشکل را حل کنیم. با این کار، میتوانیم از بار محاسباتی کمتری در مدلسازی استفاده کرده و در مقایسه با مجموعه داده­های اولیه، مدل ایجاد شده با دقت بیشتری پیش­بینی و پایداری داشته باشد .

روش رو به جلو در انتخاب ویژگی تکنیکی است برای انتخاب زیرمجموع­هایی از ویژگی­ها از داده­های اصلی که بیشترین ارتباط را برای کار پیش­بینی دارند. ایده این است که با مجموعه­ای خالی از ویژگی­ها شروع کنید و هر بار یک ویژگی را بر اساس میزان بهبود عملکرد مدل اضافه کنید. این فرآیند تا زمانی تکرار میشود که با افزودن ویژگی­های بیشتر، بهبود بیشتری حاصل نشود.



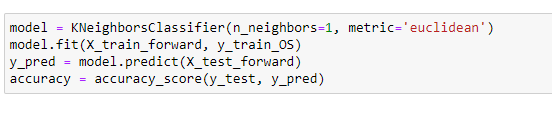
شکل 2-28 کد روش جست و جو رو به جلو

در نهایت با استفاده از این روش، ویژگی­های زیر را خواهیم داشت:

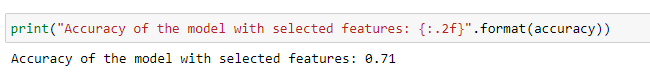


شکل 2-29 ویژگی انتخابی از روش جست و جو رو به جلو

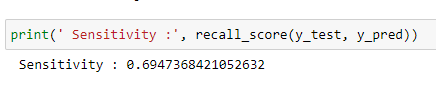
در ادامه برای اعتبار سنجی روش ارائه شده، معیارهای دقت، حساسیت و اختصاصیت را گزارش می­کنیم:



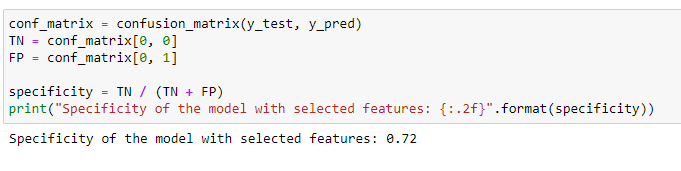
شکل 2-30 ارائه مدل برای اعتبارسنجی



شکل 2-31 گزارش دقت



شکل 2-32 گزارش حساسیت

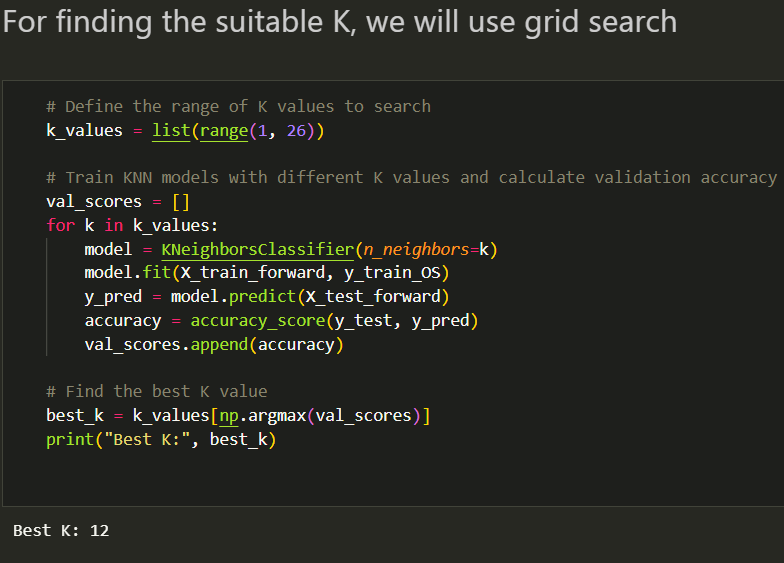


شکل 2-33 گزارش اختصاصیت

با توجه به معیارهای گزارش شده، میتوانیم نتیجه بگیریم، روش انتخاب ویژگی از سه منظر دقت، حساسیت و اختصاصیت، عملکرد مطلوب و قابل قبولی داشته است.

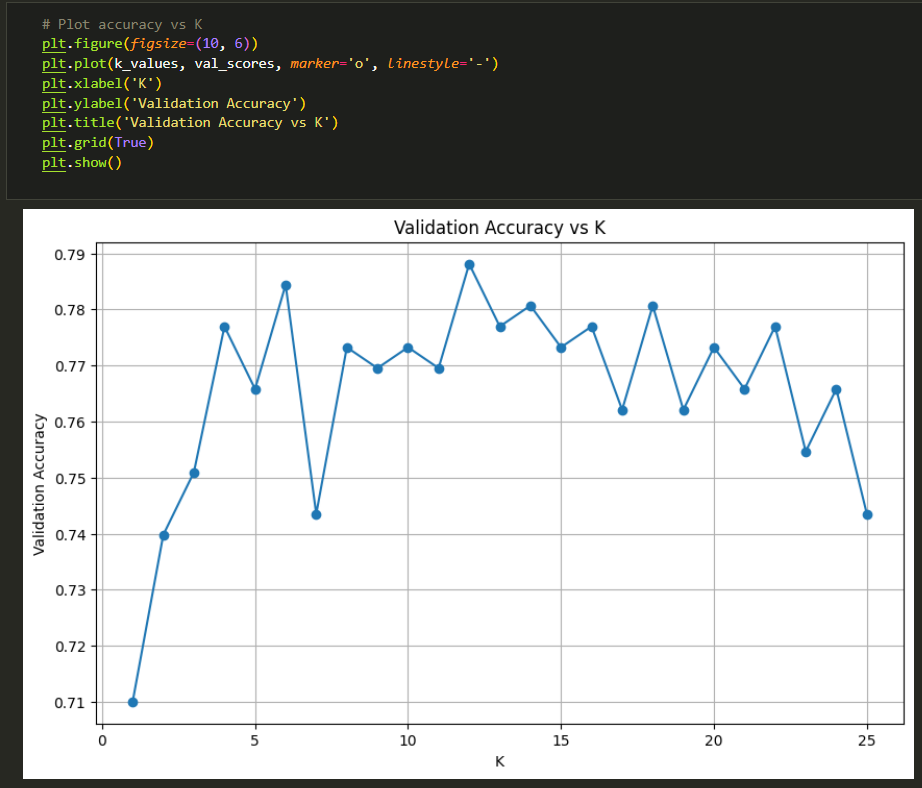
در این بخش به اجرای الگوریتم KNN بر روی دیتاست جدید می‌پردازیم. دیتاست حاضر استاندارد شده، 65 درصد داده‌ها برای آموزش جدا شده‌اند و تعداد لیبل‌های صفر و یک در آن برابر است.

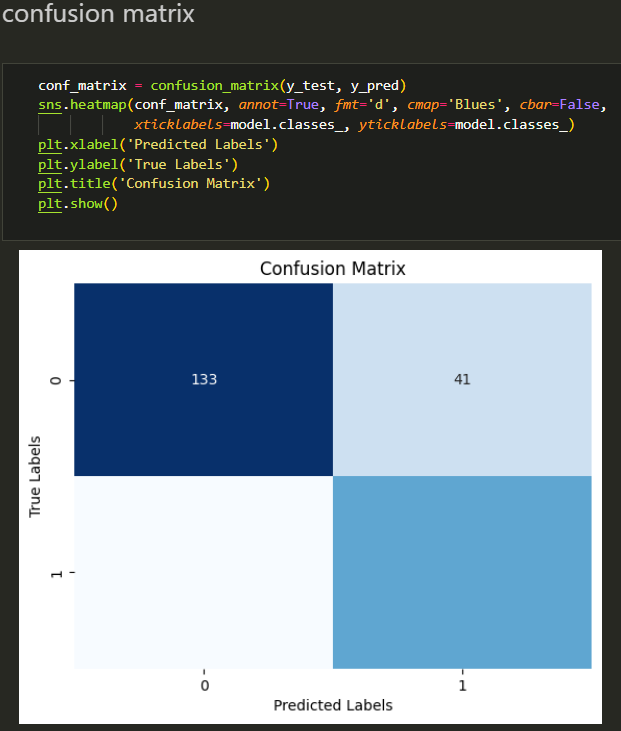
برای پیدا کردن K مناسب ابتدا محدوده‌ی مناسب برای K را تعریف می‌کنیم. حد بالای مقدار K برابر با است. 652 داده آموزشی داریم، ریشه دوم این عدد برابر با 25.53 است. بنابراین یک حد بالای مناسب برای جست و جو را می‌توان در نظر گرفت. برای پیدا کردن K مناسب از یک حلقه استفاده می‌کنیم. به طوری که ابتدا K را برابر 1 قرار می‌دهیم، مدلی را با این پارامتر آموزش می‌دهیم و تست می‌کنیم، مقدار دقت را محاسبه می‌کنیم و مقدار به دست آمده را در یک لیست ذخیره می‌کنیم. سپس با استفاده از تابع argmax() از کتابخانه نامپای، K متناظر با بیشترین مقدار دقت را نمایش می‌دهیم. شکل () کدی که وظیفه پیدا کردن K را به عهده دارد را نمایش می‌دهد.

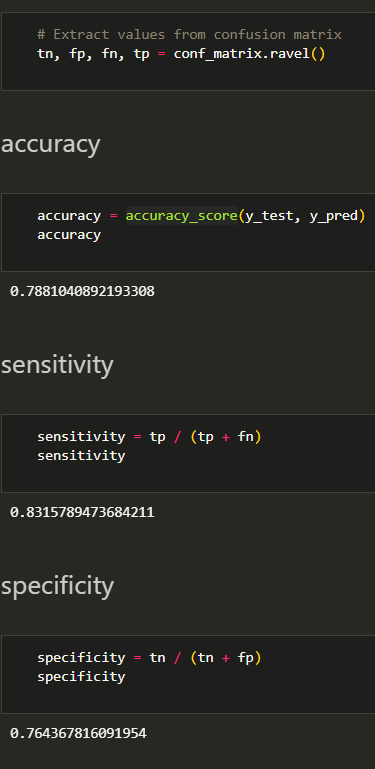


با توجه به بررسی صورت گرفته، بهترین دقت را دارد. همچنین نمودار دقت بر اساس K را نیز رسم می‌کنیم. شکل () این نمودار را نمایش می‌دهد.

حال بار دیگر با پارامتر و همین دیتاست، مدل را آموزش و ارزیابی می‌کنیم. سپس ماتریس درهم‌ریختگی، و معیارهای دقت، حساسیت و اختصاصیت را با استفاده از فرمول‌های مربوطه محاسبه می‌کنیم. برای محاسبه دقت از تابع پیش‌ساخته accuracy\_score() در سایکیت لرن استفاده می‌کنیم. شکل () ماتریس درهم ریختگی و شکل () محاسبه سه معیار را نمایش می‌دهد.

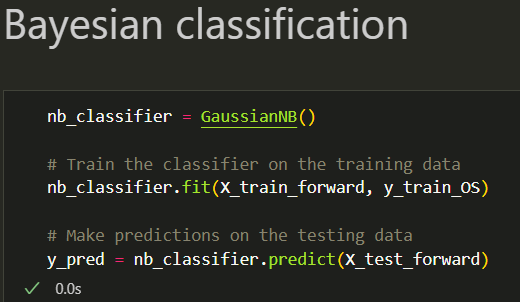




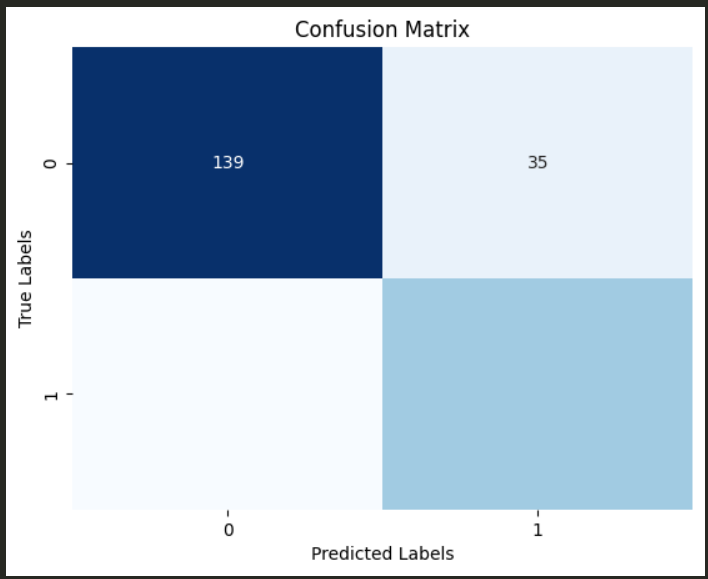


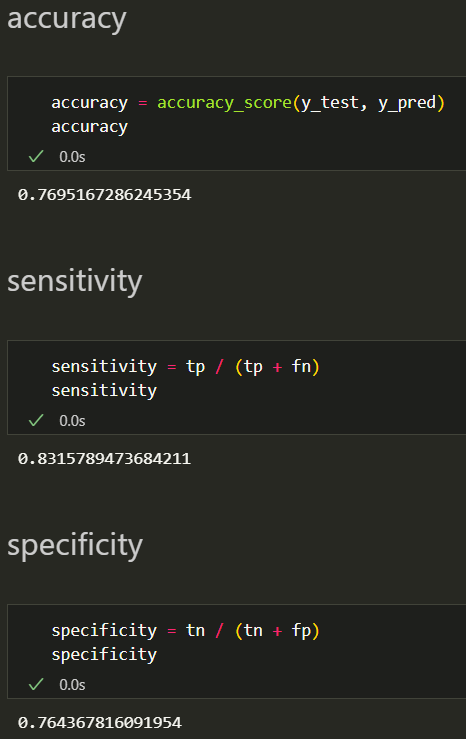
## دسته‌بندی بیزی

برای انجام دسته‌بندی بیزی از همان دیتاست بخش قبلی استفاده می‌کنیم. با استفاده از کلاس GaussianNB() از ماژول scikit-learn مدل بیزی را آموزش می‌دهیم. برای این منظور ابتدا یک شی به نام nb\_classifier از این کلاس می‌سازیم و با پاس دادن دیتاست ویژگی‌ها و ستون لیبل متناظر آنها به عنوان آرگومان به متد fit()، مدل دسته‌بندی بیزی راآموزش می‌دهیم. در ادامه با استفاده از متد predict() و پاس دادن دیتاست تست، پیش‌بینی‌های مدل از این دیتاست را به دست می‌آوریم و در متغیر y\_pred ذخیره می‌کنیم. شکل () نحوه انجام این عمل را نمایش می‌دهد.



سپس همانند دو بخش قبلی به محاسبه ماتریس درهم‌ریختگی و معیارهای ارزیابی می‌پردازیم. شکل () ماتریس درهم‌ریختگی روش دسته‌بندی بیز را نشان می‌دهد. همچنین شکل () مقدار سه معیار دقت حساسیت و اختصاصیت را برای روش دسته‌بندی بیز نشان می‌دهد.





1. Data Set [↑](#footnote-ref-1)
2. مخزن داده­های استاندارد دانشگاه کالیفرنیا [↑](#footnote-ref-2)
3. K-Nearest­\_Neighbor [↑](#footnote-ref-3)
4. Body mass index  [↑](#footnote-ref-4)
5. Comma-Separated Values [↑](#footnote-ref-5)
6. Noise [↑](#footnote-ref-6)
7. Six Sigma [↑](#footnote-ref-7)