به نام خدا

دانشگاه صنعتی امیرکبیر

دانشکده مهندسی برق

یادگیری ماشین

تمرین عملی بیزین

	زهرا لطیفی - ۹۹۲۳۰۶۹

فهرست گزارش سوالات (لطفاً پس از تکمیل گزارش، این فهرست را به‌روز کنید.)

سوال ۱ – تشخیص دیابت

در این سوال قرار است که افرادی که دیابت دارند را تشخیص دهید. به این لینک برای دریافت دیتاست بروید. این دیتاست ۸ ویژگی به همراه ۲ کلاس دارد. آنالیز بیزین برروی این دیتاست باید انجام شود.

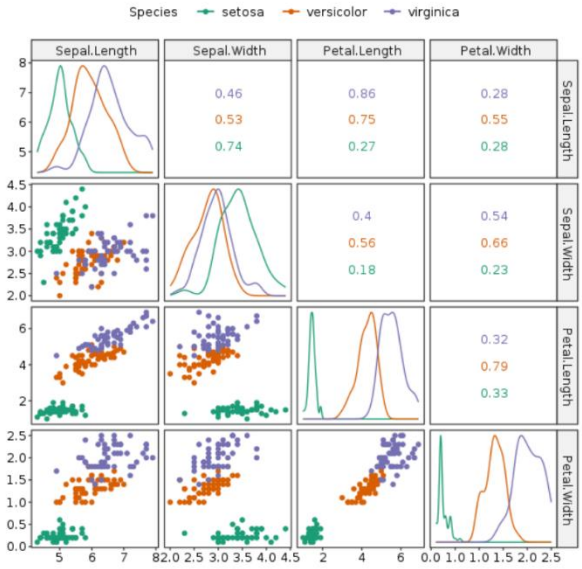
۱- با استفاده از اسکترپلات رابطه بین فیچر های مختلف را نشان دهید. (همانند شکل ۱ اپندیکس)

۲- دیتاست را به صورت ۷۰ به ۳۰ برای آموزش و تست جدا کنید.

۳- مدل سازی : با استفاده از Gaussian Naïve Bayes دسته بندی را انجام دهید. (با استفاده از کتابخانه)

۴- ماتریس آشفتگی را برای کلاسیفایر های روی آموزش و تست نشان دهید.

Appendix :



گزارش سوال ۱

ابتدا کتابخانه‌های numpy، pandas، seaborn، matplotlib، sklearn را اضافه کردیم. سپس دیتاست مربوط به diabetes_prediction_dataset را بارگذاری کرده و ۱۰ ردیف اول آن را نمایش دادیم. مشاهده کردیم که برخی از فیچرها Categorical بوده و برخی هم نرمالایز نشده بودند. پس در گام اول، داده‌های مربوط به فیچر smoking_history و gender را با تابع preprocessing.LabelEncoder() انکد کردیم.

```
# Create an instance of the encoder & fit it
label_encoder = preprocessing.LabelEncoder()
label_encoder.fit(dataset['smoking_history'])
```

```
dataset['smoking_history'] =
label_encoder.transform(dataset['smoking_history'])
label_encoder.fit(dataset['gender'])
dataset['gender'] = label_encoder.transform(dataset['gender'])
```

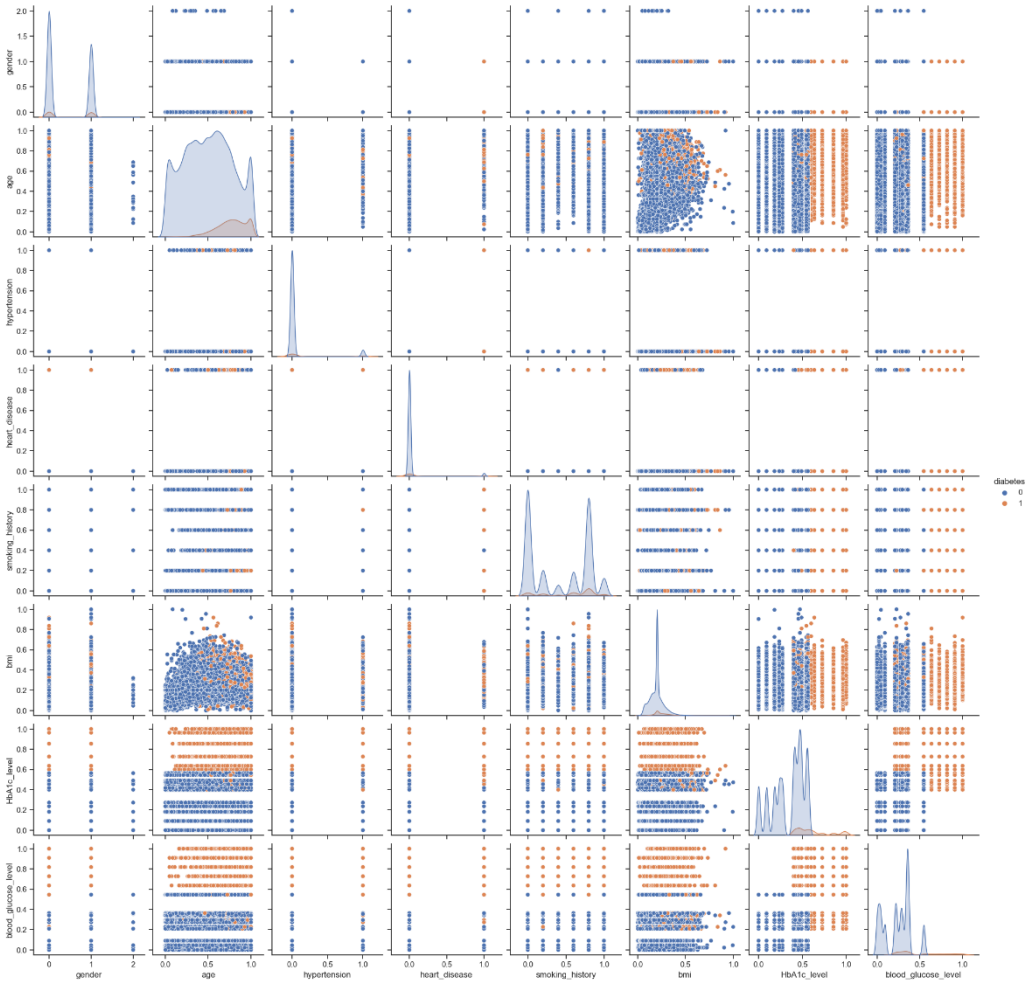
سپس با تکه کد زیر، داده‌های نرمالایز نشده را بین ۰ و ۱ نرمالایز کردیم:

```
# Apply normalization techniques [0, 1]
for column in ['age','bmi', 'HbA1c_level', 'blood_glucose_level',
'smoking_history']:
    dataset[column] = (dataset[column] - dataset[column].min()) /
(dataset[column].max() - dataset[column].min())
```

و بار دیگر ۱۰ ردیف اول دیتاست را نمایش دادیم.

۲) داده‌های ستون diabetes که تارگت ماست را به عنوان y و سایر ستون‌ها که فیچر هستند را به عنوان x داده ذخیره می‌کنیم. نهایتا با دستور train_test_split داده‌های تست و آموزش را با نسبت ۷۰/۳۰ جدا می‌کنیم.

۱) برای رسم اسکترپلات بین فیچرها، از کتابخانه seaborn و تابع sns.pairplot استفاده کردیم.



۳) برای دسته‌بندی با Gaussian Naïve Bayes، NB.GaussianNB() را فراخوانی کرده، به داده‌های آموزش خود، فیت کرده، دقت را بر روی هر دوره ولیدیشن گزارش کردیم:

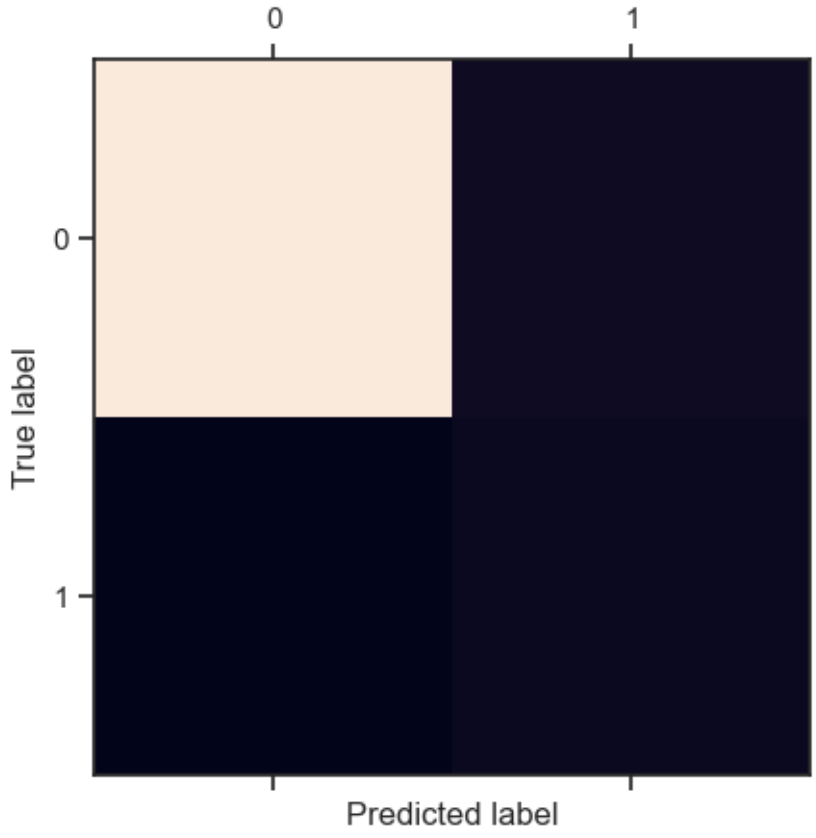
```
Accuracy of each validation:
[0.90614286 0.90042857 0.90514286 0.90271429 0.903      0.90485714
 0.90671429 0.90357143 0.90314286 0.90442857]
```

پس از آن، بر داده‌های تست اعمال کرده و دقت نهایی را گزارش کردیم:

```
Total accuracy:
          0.9059666666666667
```

۴) ماتریس آشفتگی را برای داده‌های آموزش محاسبه و رسم کردیم:

```
[[25536  1917]
 [  904  1643]]
True Negatives(TN) =  25536
True Positives(TP) =  1643
False Positives(FP) =  1917
False Negatives(FN) =  904
```

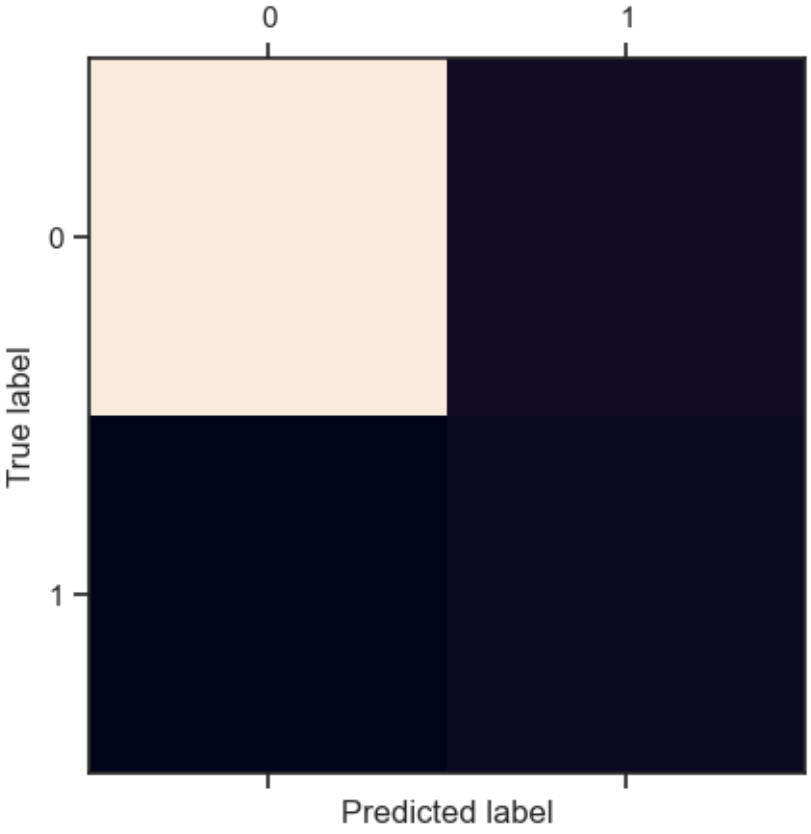


بار دیگر همین روند را برای داده‌های آموزش انجام دادیم تا ماتریس آشفتگی را برای داده‌های آموزش هم رسم کنیم:

```
Total accuracy:
0.9039428571428572
```

```
[[59403  4644]
 [ 2080  3873]]
True Negatives(TN) =  59403
True Positives(TP) =  3873
False Positives(FP) =  4644
False Negatives(FN) =  2080
```

نتایج کمی بهبود یافت که انتظارش را داشتیم.



نکته حائز اهمیت درباره این تمرین، این است که دیتاست بالانس نبود. تعداد داده‌هایی که در کلاس دارای دیابت (۱) بودند، ۹۱۵۰۰ و تعداد داده‌های کلاس (۰)، ۸۵۰۰ عدد بود. این اختلاف باعث شد که در نتیجه، علی‌رغم اینکه True Negativeها بسیار بیشتر از False Negativeها شدند، برای داده‌های Positive این اتفاق نیفتاده و Falseها از Trueها بیشتر شوند. راه‌حل، اضافه کردن داده به دیتاست و یا Data Augmentation است.

سوال ۲ – تشخیص سرطان سینه

مجموعه داده ای در فایل پیوست قرار دارد، شامل مجموعه ای از ویژگی های اندازه گیری شده پزشکی برای تشخیص سرطان سینه. داده ها در فایل BreastCancer.csv قرار داده شده اند. بتدا داده ها را به صورت یکنواخت از هر دو کلاس بین داده های آموزش، ارزیابی و آزمون تقسیم بندی کنید.

الف) ابتدا با استفاده از مدل Naïve Bayes و با روش Cross validation عملکرد این الگوریتم را بر روی داده ها مشاهده نمایید و Accuracy و ماتریس درهم ریختگی را گزارش کنید.

ب) این بار با استفاده از مدل Gaussian Naïve Bayes فرایند فوق را تکرار کنید.

گزارش سوال ۲

الف) ابتدا کتابخانه‌های numpy، pandas، seaborn، matplotlib، sklearn را اضافه کردیم. سپس دیتاست مربوط به BreastCancer را بارگذاری کرده و ۱۰ ردیف اول آن را نمایش دادیم. مشاهده کردیم که برخی از فیچرها نرمالایز نشده بودند. پس در گام اول، همانند سوال اول داده‌های نرمالایز نشده را بین ۰ و ۱ نرمالایز کردیم. با دستور split_test_trainداده‌ها را به نسبت ۰.۱ از کل داده‌ها به آموزش و تست تقسیم کردیم. همچنین روی داده ها شافل اجرا کردیم تا به صورت رندم جابجا شوند و در نهایت ده ردیف ابتدایی دیتاست را نمایش دادیم.

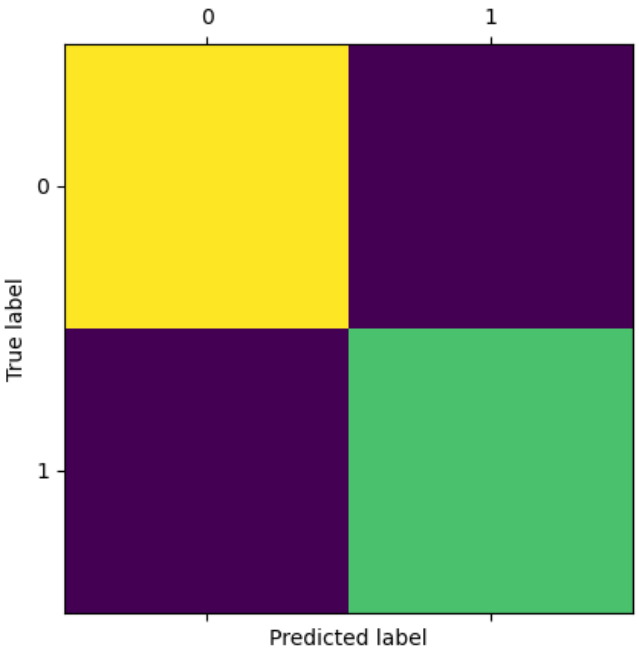
در مرحله بعد باید مدل بیزین مورد نظر را مشخص کنیم. به دلیل اینکه داده‌ها به صورت گسسته مقداردهی شده‌اند، از Multi_naive_bayesاستفاده می‌کنیم. پس از معرفی مدل به کمک تابع fit مدل را آموزش می‌دهیم. برای این کار از تمامی داده‌های train استفاده می‌کنیم. در نهایت مدل آموزش داده شده را به روش k_fold_cross_validation اصلاح می‌کنیم. ورودی‌های تابع cross_val_score مدل آموزش داده شده، تمامی داده‌های train و تعداد دسته‌های مورد نیاز را ۱۰ انتخاب می‌کنیم:

```
Accuracy of each validation:
[0.94949495 0.96938776 0.92857143 0.97959184 0.91836735]
```

در ادامه score بدست آمده یا همان accuracy را محاسبه می‌کنیم و در نهایت نیز ماتریس آشفتگی را نمایش می‌دهیم.

```
Total accuracy:
0.9562043795620438
[[76  3]
 [ 3 55]]
True Negatives(TN) =  76
True Positives(TP) =  55
False Positives(FP) =  3
False Negatives(FN) =  3
```

همانطور که مشخص است در این جا با دقت خوبی مدل ترین شده است. درایه ۰،۰ این ماتریس نشان دهنده تعداد کسانی است که مبتلا به سرطان نیستند و به درستی تشخیص داده شده، درایه ۱،۱ این ماتریس نشان دهنده مبتلایان به سرطان است که به درستی تشخیص داده شده‌اند و درایه ۰،۱ مربوط به مبتلایان به سرطان است که اشتباه تشخیص داده شده و در نهایت دیگر درایه، مربوط به کسانی است که مبتلا به سرطان نیستند و اشتباه تشخیص داده شده.



ب) درست مانند سوال قبل و بخش قبل همین سوال Gaussian Naïve Bayes را هم آموزش دادیم که نتایج آن به شرح زیرند:

```
accuracy of each validation:
[0.84848485 0.87755102 0.7755102  0.86734694 0.84693878]

total accuracy:
0.8029197080291971

[[79  0]
```

```
[27 31]]
True Negatives(TN) =  79
True Positives(TP) =  31
False Positives(FP) =  0
False Negatives(FN) =  27
```

همانطور که انتظار داشتیم، خروجی این روش بهتر از روش قبل شد.