# «به نام خدا»

## تکلیف چهارم – سوال دوم – مرضیه علیدادی – 9631983

(کد های مربوط، در دو فرمت py. و ipynb. ضمیمه شده اند.)

---

## .2

**a)**

| | Number of times pregnant. | Plasma glucose concentration a 2 hours in an oral glucose tolerance test. | Diastolic blood pressure (mm Hg). | Triceps skinfold thickness (mm). | 2-Hour serum insulin (mu U/ml). | Body mass index (weight in kg/(height in m)^2). | Diabetes pedigree function. | Age (years). | Class variable (0 or 1). |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 6 | 148 | 72 | 35 | 0 | 33.6 | 0.627 | 50 | 1 |
| **1** | 1 | 85 | 66 | 29 | 0 | 26.6 | 0.351 | 31 | 0 |
| **2** | 8 | 183 | 64 | 0 | 0 | 23.3 | 0.672 | 32 | 1 |
| **3** | 1 | 89 | 66 | 23 | 94 | 28.1 | 0.167 | 21 | 0 |
| **4** | 0 | 137 | 40 | 35 | 168 | 43.1 | 2.288 | 33 | 1 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **763** | 10 | 101 | 76 | 48 | 180 | 32.9 | 0.171 | 63 | 0 |
| **764** | 2 | 122 | 70 | 27 | 0 | 36.8 | 0.340 | 27 | 0 |
| **765** | 5 | 121 | 72 | 23 | 112 | 26.2 | 0.245 | 30 | 0 |
| **766** | 1 | 126 | 60 | 0 | 0 | 30.1 | 0.349 | 47 | 1 |
| **767** | 1 | 93 | 70 | 31 | 0 | 30.4 | 0.315 | 23 | 0 |

768 rows × 9 columns

**b)** برای حذف missing value ها، در بین فیلد ها، Null وجود ندارد. ولی Attribute های 2ام و 3ام و 4ام و 5ام و 6ام دارای مقدار 0 در برخی سطر ها هستند، که این نشاندهنده ی مقدار نامعتبر است. پس این ها را با میانگین ستون جایگزین کردم. درنهایت، رنج متغیر ها به این صورت شد:

| | Number of times pregnant. | Plasma glucose concentration a 2 hours in an oral glucose tolerance test. | Diastolic blood pressure (mm Hg). | Triceps skinfold thickness (mm). | 2-Hour serum insulin (mu U/ml). | Body mass index (weight in kg/(height in m)^2). | Diabetes pedigree function. | Age (years). | Class variable (0 or 1). |
|---|---|---|---|---|---|---|---|---|---|
| **min** | 0 | 44.0 | 24.0 | 7.0 | 14.0 | 18.2 | 0.078 | 21 | 0 |
| **max** | 17 | 199.0 | 122.0 | 99.0 | 846.0 | 67.1 | 2.420 | 81 | 1 |

داده ها پس از نرمال سازی داده های عددی:

| | Number of times pregnant. | Plasma glucose concentration a 2 hours in an oral glucose tolerance test. | Diastolic blood pressure (mm Hg). | Triceps skinfold thickness (mm). | 2-Hour serum insulin (mu U/ml). | Body mass index (weight in kg/(height in m)^2). | Diabetes pedigree function. | Age (years). | Class variable (0 or 1). |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 0.025315 | 0.624447 | 0.303785 | 0.147673 | 0.656295 | 0.141766 | 0.002645 | 0.210962 | 1 |
| **1** | 0.005111 | 0.434404 | 0.337302 | 0.148208 | 0.794950 | 0.135943 | 0.001794 | 0.158430 | 0 |
| **2** | 0.031558 | 0.721898 | 0.252467 | 0.115004 | 0.613606 | 0.091914 | 0.002651 | 0.126233 | 1 |
| **3** | 0.006612 | 0.588467 | 0.436392 | 0.152076 | 0.621527 | 0.185797 | 0.001104 | 0.138852 | 0 |
| **4** | 0.000000 | 0.596386 | 0.174127 | 0.152361 | 0.731335 | 0.187622 | 0.009960 | 0.143655 | 1 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **763** | 0.042321 | 0.427443 | 0.321640 | 0.203141 | 0.761779 | 0.139236 | 0.000724 | 0.266623 | 0 |
| **764** | 0.009245 | 0.563972 | 0.323590 | 0.124813 | 0.719056 | 0.170116 | 0.001572 | 0.124813 | 0 |
| **765** | 0.026915 | 0.651352 | 0.387582 | 0.123811 | 0.602905 | 0.141037 | 0.001319 | 0.161492 | 0 |
| **766** | 0.004582 | 0.577294 | 0.274902 | 0.133572 | 0.712675 | 0.137909 | 0.001599 | 0.215340 | 1 |
| **767** | 0.004990 | 0.464076 | 0.349304 | 0.154692 | 0.776195 | 0.151698 | 0.001572 | 0.114771 | 0 |

768 rows × 9 columns

تنها متغیر دسته ای، 'Class variable (0 or 1).' است، که نیازی به اینکد کردن ندارد و دارای دو دسته 0 و 1 است.

**d)** توزیع دسته های ستون 'Class variable (0 or 1).' به ترتیب در دو مجموعه ی آموزش و تست:

```
y_train.value_counts()
```

```
0    391
1    223
Name: Class variable (0 or 1)., dtype: int64
```

```
y_test.value_counts()
```

```
0    109
1     45
Name: Class variable (0 or 1)., dtype: int64
```

**f)** با استفاده از قابلیت های sklearn، با استفاده از مدل تولید شده، متغیر هدف را برای داده های تست پیشبینی کردم. نتیجه را با مقادیر واقعی متغیر هدف مقایسه کردم. بدین صورت شد:

```
print(confusion_matrix(y_test, y_pred))
```

```
[[101   8]
 [ 37   8]]
```

```
print(classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       0.73      0.93      0.82       109
           1       0.50      0.18      0.26        45

    accuracy                           0.71       154
   macro avg       0.62      0.55      0.54       154
weighted avg       0.66      0.71      0.66       154
```

- ماتریس تهیه شده بدین صورت تفسیر می شود:

هر سطر نشان دهنده ی یکی از دسته های واقعی (actual) است. و هر ستون نشاندهنده ی یکی از دسته های پیشبینی شده.

داده هایی که در دسته ی 0 قرار می گیرند، 109 تا هستند. که 101 مورد از آن ها درست پیشبینی شده اند.

داده هایی که در دسته ی 1 قرار می گیرند، 45 تا هستند. که 8 تا از آن ها درست پیشبینی شده اند.

- گزارشی که در ادامه آمده، همین مطالب را به صورت درصدی بیان می کند:

دقت پیشبینی داده های دسته های 0 و 1، به ترتیب برابر 93% و 18% است.

و به طور کلی، دقت مدل برابر 71% است.

**g)** مدل را با max_depth های مختلف ایجاد کردم. تا max_depth=12 دقت به 73% رسید. پس از آن، با افزایش max_depth، دقتی که حاصل می شد، کمتر مساوی این مقدار بود. پس max_depth=12 بهترین حالت است.