

تمرین شماره ۱

نیمسال اول ۴۰۲

داده‌های مربوط به پیش بینی شروع بیماری قند در زیر آمده است.

The **Pima Indians Diabetes Dataset** involves predicting the onset of diabetes within 5 years in Pima Indians given medical details.

It is a binary (2-class) classification problem. The number of observations for each class is not balanced. There are 768 observations with 8 input variables and 1 output variable. Missing values are believed to be encoded with zero values. The variable names are as follows:

1. Number of times pregnant.
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test.
3. Diastolic blood pressure (mm Hg).
4. Triceps skinfold thickness (mm).
5. 2-Hour serum insulin (mu U/ml).
6. Body mass index (weight in kg/(height in m)²).
7. Diabetes pedigree function.
8. Age (years).
9. Class variable (0 or 1).

Link:

<https://raw.githubusercontent.com/jbrownlee/Datasets/master/pima-indians-diabetes.csv>

الف) با استفاده از پایتون برای هر ویژگی، هیستوگرام آن را رسم نمایید (بدون توجه به کلاس داده).

ب) هیستوگرام ویژگی‌های شماره ۲، ۳، ۶ و ۷ را با توجه به تعلق هر داده به کلاس رسم نمایید. (چهار نمودار رسم شود که هر یک دربردارنده دو هیستوگرام مربوط به یک ویژگی برای دو کلاس مجزا است).

ج) اگر ویژگی ۶ دارای توزیع احتمال نرمال باشد. میانگین و واریانس آن را محاسبه کرده و در یک صفحه، تابع چگالی احتمال و هیستوگرام داده‌ها را باهم رسم نمایید.