تمرین انتخاب ویژگی

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

۱) الف) ویژگی ۳ و ۶ مربوط به فشارخون و توده بدنی است. نمودار ROC را برای این دو رسم به صورت جداگانه نمایید (حل با ابزار).

ب) با توجه به LDA داده‌های بر روی یک خط با بیشترین جداکنندگی تصویر نمایید و پراکندگی آنها را بر روی این خط بررسی نمایید. یک نقطه آستانه مناسب برای جداسازی داده‌ها ارایه دهید.

* برای آنکه نتایج بهتری به دست آید. نمونه‌هایی که دارای مقدار صفر هستند که نشانه بی‌مقدار بودن آن کمیت (نویز) است را معین و حذف نمایید.

۲) با توجه به LDA خط با بیشترین جداکنندگی را برای داده‌های زیر بیابید