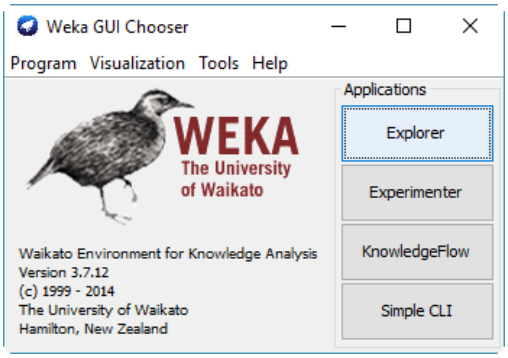
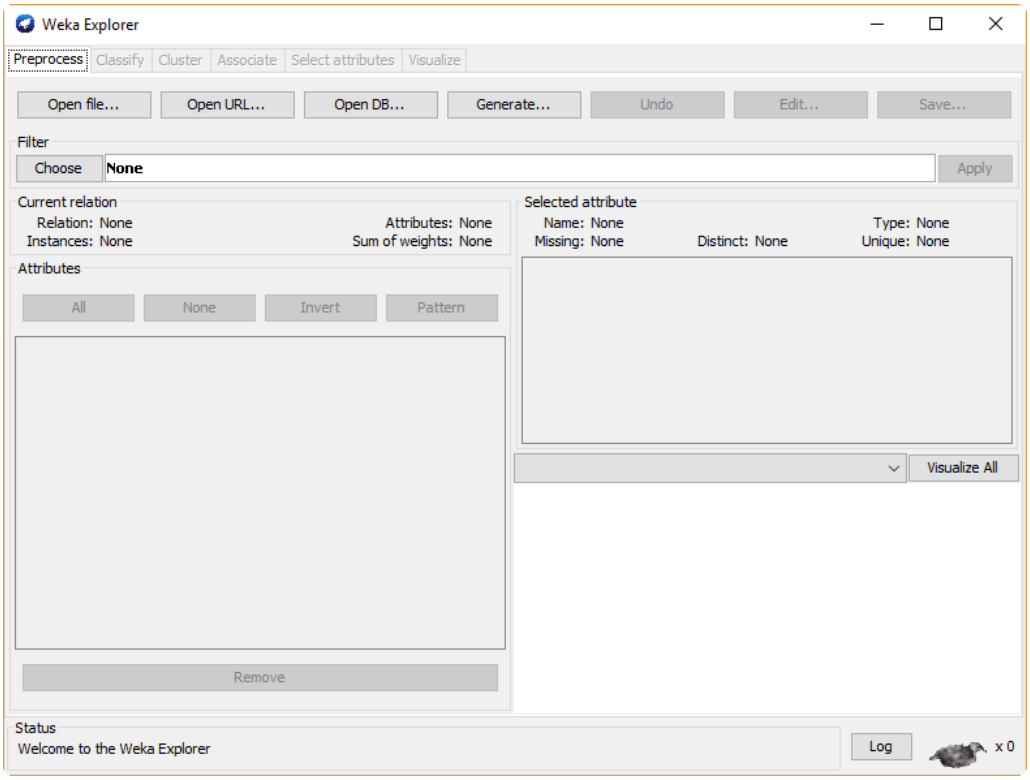
به نام او ...

تمرین سوم درس داده کاوی

هدف از این تمرین آشنایی با ابزار Weka می‌باشد. Weka را می‌توانید از این لینک دریافت کنید. پس از اجرای نرم افزار وارد قسمت Explorer شوید.



در تب Preprocess می‌توانید از بخش Open file مجموعه داده‌ی مورد نظر را بارگذاری کنید.



مجموعه داده‌های مورد استفاده در Weka معمولا به فرمت arff می‌باشند. این فرمت مانند فرمت csv می‌باشد با این تفاوت که در بخش header فایل مثل شکل زیر اطلاعاتی مانند نام و نوع ویژگی‌ها نیز آورده می شود.

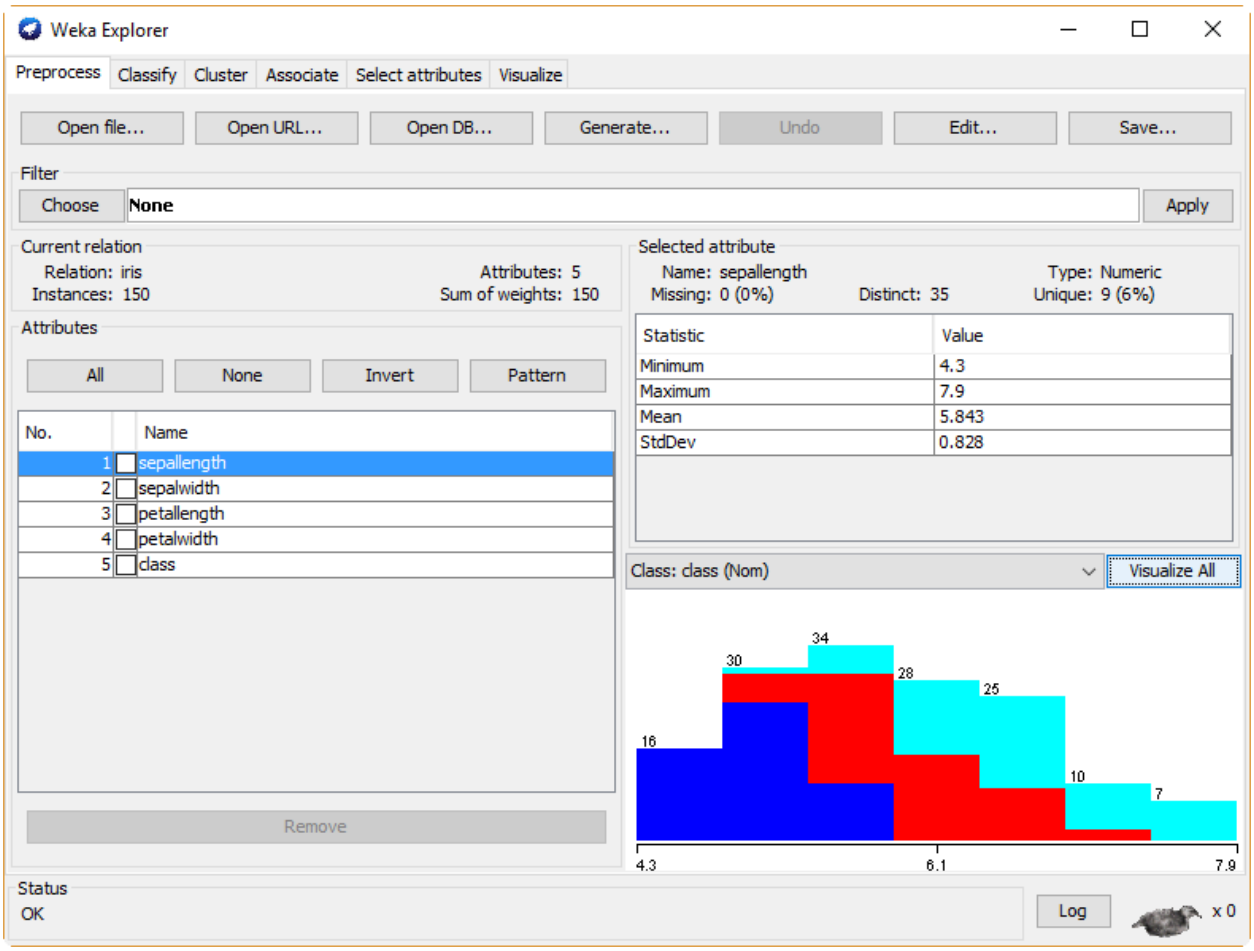
```
@relation 'labor-neg-data'
@attribute 'duration' numeric
@attribute 'wage-increase-first-year' numeric
@attribute 'wage-increase-second-year' numeric
@attribute 'wage-increase-third-year' numeric
@attribute 'cost-of-living-adjustment' {'none','tcf','tc'}
@attribute 'working-hours' numeric
@attribute 'pension' {'none','ret_allw','empl_contr'}
@attribute 'standby-pay' numeric
@attribute 'shift-differential' numeric
@attribute 'education-allowance' {'yes','no'}
@attribute 'statutory-holidays' numeric
@attribute 'vacation' {'below_average','average','generous'}
@attribute 'longterm-disability-assistance' {'yes','no'}
@attribute 'contribution-to-dental-plan' {'none','half','full'}
@attribute 'bereavement-assistance' {'yes','no'}
@attribute 'contribution-to-health-plan' {'none','half','full'}
@attribute 'class' {'bad','good'}
@data
1,5,?,?,?,40,?,?,2,?,11,'average',?,?,'yes',?,'good'
2,4.5,5.8,?,?,35,'ret_allw',?,?,'yes',11,'below_average',?,'full',?,'full','good'
?,?,?,?,?,38,'empl_contr',?,5,?,11,'generous','yes','half','yes','half','good'
```

۱- قسمت header مورد نیاز برای فایل sample را با توجه به فایل sample_description تولید کرده و با فرمت arff ذخیره کنید. برای بررسی صحت کار خود فایل تغییر داده شده را در Weka بارگذاری کنید. این فایل را در پوشه‌ی نهایی خود برای بارگذاری در سایت درس قرار دهید.

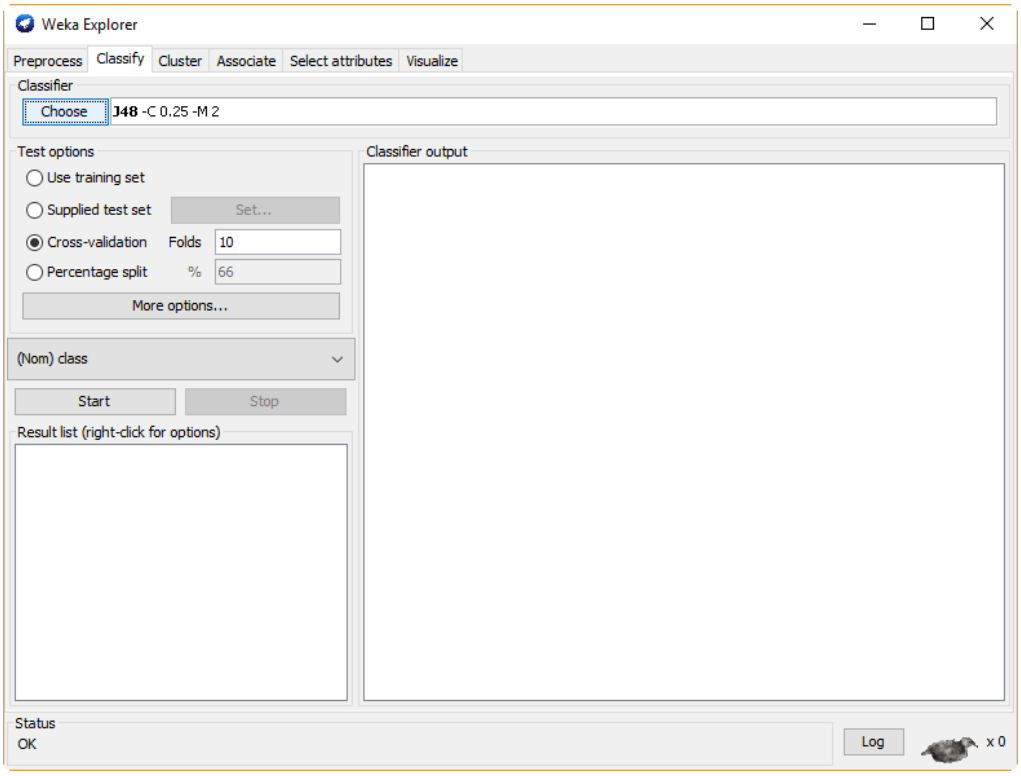
در ادامه مجموعه داده‌ی Iris را از پوشه‌ی data بارگذاری کنید. همانطور که مشاهده می‌کنید اطلاعات کلی از مجموعه داده پس از بارگذاری نشان داده می‌شود. از قسمت Edit می‌توانید مقادیر مجموعه داده را در جدولی مشاهده کرده و تغییرات دلخواه خود را در مقادیر آن اعمال کنید. در قسمت Attributes می‌توانید ویژگی مورد نظر خود را انتخاب کرده و در قسمت Selected attribute توضیحات مربوط به این ویژگی را مشاهده کنید.

۲- از قسمت Visualize All نمودار پراکندگی سه کلاس به ازای هر ویژگی را نمایش داده و تحلیل کنید.

۳- از بخش Filter می‌توانید پیش پردازش‌های مختلفی را بر روی مجموعه داده انجام دهید. در این بخش به قسمت unsupervised و سپس attribute رفته و برای نرمالسازی و گسسته سازی مجموعه داده از Normalize و Descritize استفاده کنید. با کلیک بر روی فیلتر انتخاب شده می‌توانید پارامترهای مختلف آن را تغییر دهید. از این طریق فقط دو ویژگی ۳ و ۴ را با ۵ مقدار مختلف گسسته سازی کرده و در انتها تغییرات خود را از قسمت Save ذخیره کنید. فایل ذخیره شده را در پوشه‌ی نهایی قرار دهید.



در مرحله‌ی بعد تب Classify را مورد بررسی قرار می‌دهیم. مجموعه داده‌ی Iris را دوباره بارگذاری کنید. از قسمت Classifier می‌توانید الگوریتم دسته بندی را انتخاب کنید. در بخش Test options روش ارزیابی الگوریتم انتخاب می‌شود. در این جا از 10-fold Cross-Validation استفاده خواهیم کرد. پس از انتخاب الگوریتم دسته بندی متغیر هدف مورد نظر برای دسته بندی نیز در قسمت زیر Test options قابل انتخاب می‌باشد. در مجموعه داده‌ی Iris ستون class متغیر هدف می‌باشد. پس از انتخاب متغیر هدف با زدن Start فرآیند آموزش مدل شروع می‌شود. پس از آموزش، اطلاعاتی از قبیل زمان آموزش مدل، دقت به دست آمده توسط روش ارزیابی انتخاب شده و اطلاعات دیگری مانند TPR، FPR و Confusion Matrix نمایش داده می‌شود.



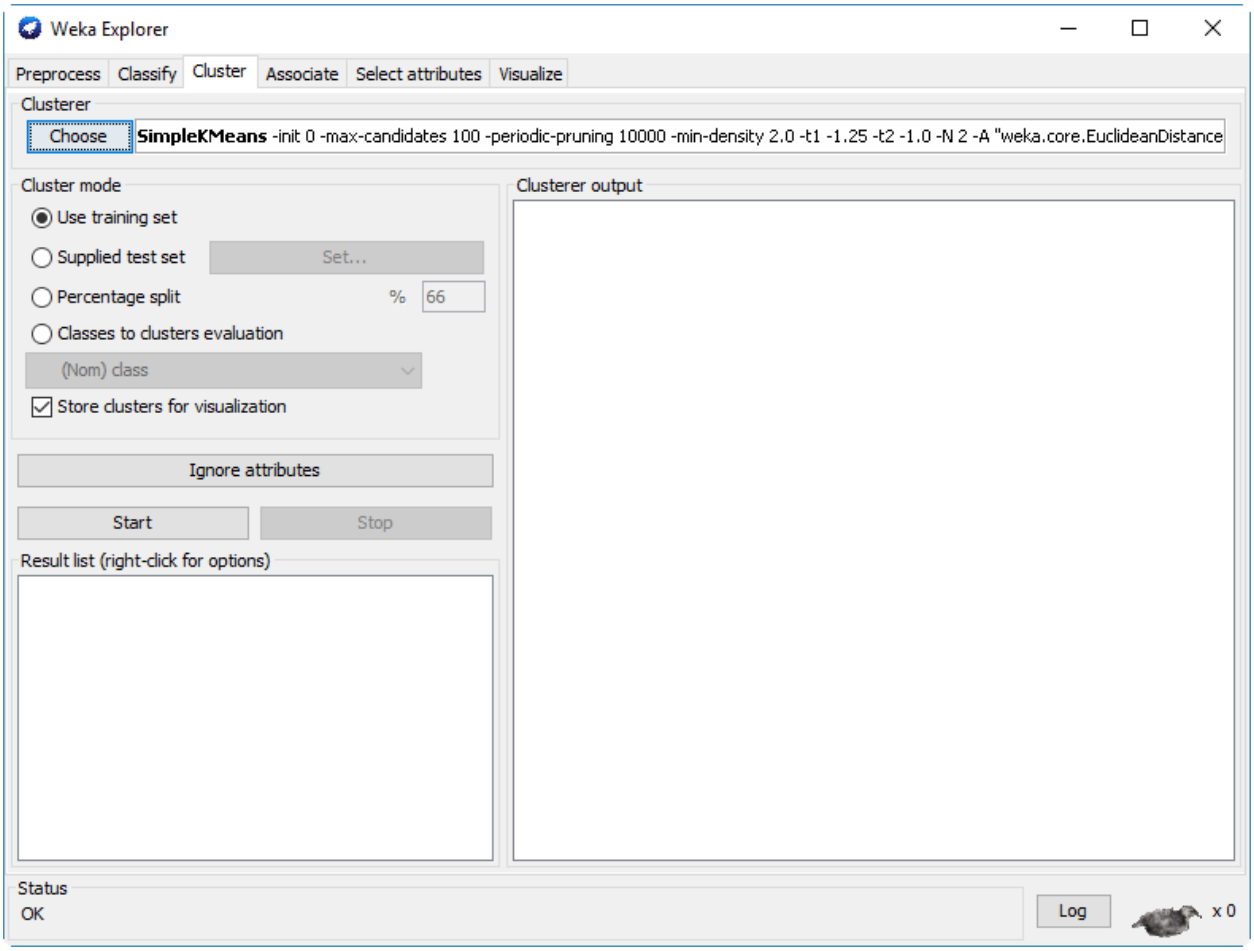
۴- در بین روش‌های دسته بندی از قسمت lazy روش IBk که همان KNN می‌باشد را انتخاب کنید. با مقادیر K=1, 5, 15, 25, 50, 100, 150 دقت روش را بر روی مجموعه داده‌ی Iris به دست آورده، نمودار دقت بر حسب مقدار K را رسم کرده و آن را تحلیل کنید.

۵- از قسمت trees دو روش J48(درخت تصمیم) و Decision Stump را انتخاب کرده و بر روی مجموعه داده‌ی Iris آموزش دهید. دقت، TPR، FPR، Precision و Confusion Matrix این دو روش را با هم مقایسه کنید.

در صورتی که با یک مسئله‌ی رگرسیون مواجه باشید از طریق تب Classify همانند یک مسئله‌ی دسته بندی می‌توانید روش دلخواه خود را انتخاب کنید. با توجه به نوع متغیر هدف انتخاب شده، Weka نوع مسئله (دسته بندی یا رگرسیون) را تشخیص داده و فقط روش‌های موجود برای آن مسئله‌ی خاص قابل انتخاب خواهند بود.

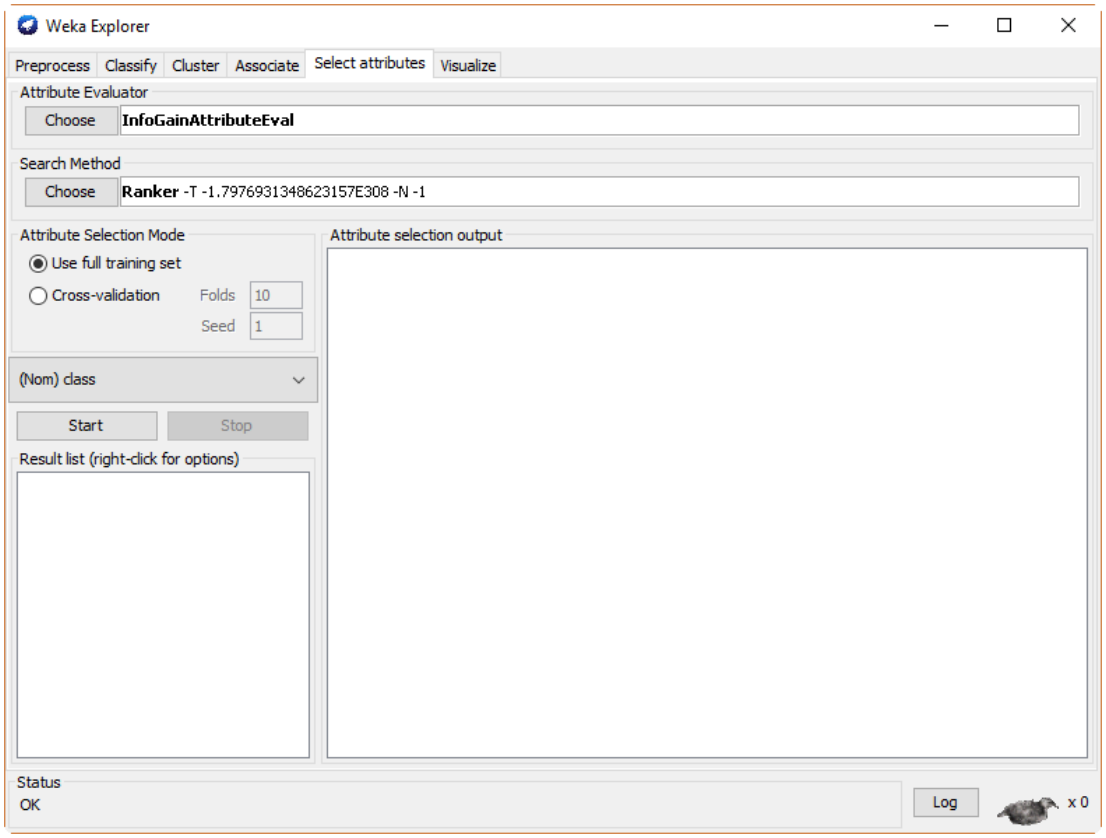
۶- از تب Preprocess متغیر class را از مجموعه داده‌ی Iris حذف کنید. سپس از طریق تب Classify متغیر petalwidth را به عنوان متغیر هدف انتخاب کنید. از قسمت Classifier از بخش functions روش LinearRegression را انتخاب کنید. مدل را آموزش داده و تساوی به دست آمده برای محاسبه‌ی petalwidth را از روی سه ویژگی دیگر با توجه به وزن‌های به دست آمده گزارش کنید. RMSE مدل را نیز گزارش کنید.

در تب Cluster می‌توانید الگوریتم‌های خوشه بندی مختلف را مورد استفاده قرار دهید. (از آنجایی که ممکن است هیچ آشنایی با مبحث خوشه بندی نداشته باشید، این قسمت صرفا جنبه‌ی آموزشی برای Weka داشته و نیازی به پاسخگویی در گزارش نمی‌باشد).



در ادامه‌ی بخش ۶ تمرین و پس از حذف متغیر کلاس از قسمت Clusterer الگوریتم SimpleKmeans را انتخاب کرده و سپس پارامتر numClusters آن را برابر ۳ قرار دهید و اجرا کنید. از قسمت Result list بر روی مدل به دست آمده کلیک راست کرده و visualize cluster assignments را بزنید. برای محور Y ویژگی petalwidth و برای محور X ویژگی petallength را انتخاب کنید. تصویر نمودار به دست آمده را ذخیره کنید. دقت داشته باشید که در این مرحله برچسب کلاس‌ها در اختیار نبوده و با الگوریتم خوشه بندی برچسب‌هایی برای داده ها در نظر گرفته شد. حال مجموعه داده‌ی اصلی Iris را دوباره بارگذاری کنید تا متغیر class را نیز داشته باشید. در تب Visualize نمودار تمام ویژگی‌ها نسبت به هم موجود می‌باشد. بر روی نموداری که محور Y آن petalwidth و محور X آن petallength می‌باشد کلیک کنید. تصویر نمودار حاصل را ذخیره کنید. شباهت این دو نمودار را برای خودتان بررسی کنید.

در انتها تب Select attributes مورد بررسی قرار می‌گیرد. در این قسمت می‌توان ارزش ویژگی‌ها را با توجه به روش‌های ارزیابی مختلف مورد بررسی قرار داد.



۷- از قسمت Attribute Evaluator روش InfoGainAttributeEval را انتخاب کنید. Search Method را نیز برابر Ranker قرار دهید و متغیر numToSelect آن را برابر ۲ قرار دهید. پس از اجرا دو متغیر با بیش ترین امتیاز نمایش داده خواهند شد. این دو متغیر به همراه امتیاز آن‌ها را گزارش کنید. همین کار را برای روش CorrelationAttributeEval تکرار کنید.

گزارش:

گزارش بایستی در قالب فایل PDF باشد. لطفا فایل Word نفرستید.

فایل گزارش خود را به شکل «Report3_StdNum.pdf» نامگذاری کنید. (مانند Report3_9131081.pdf)

بارگذاری:

تمام فایل های موجود را در قالب یک فایل فشرده در سایت درس بارگذاری نمایید.

فایل فشرده را به شکل «DM3_StdNum» نامگذاری کنید. (مانند DM3_9131081)

مهلت ارسال تمرین ساعت ۲۳:۵۵ دقیقه‌ی روز جمعه مورخ ۱۱ فروردین می‌باشد.

به ازای هر روز تاخیر در ارسال تمرین، برای دو روز اول **۵** و روز های بعد **۱۰** درصد از نمره‌ی آن از دست خواهد رفت.

هر گونه سوال در مورد تمرین را می‌توانید از طریق ایمیل AUT.DM2017@gmail.com بپرسید.

موفق باشید