تمرین سری دوم داده کاوی

در این سری تمرین شما بایستی با استفاده از داده های پیوست شده و روشهای معرفی شده در اسلایدهای درخت تصمیم گیری دو مدل طراحی کنید تا درباره ابتلا یا عدم ابتلای افراد به بیماری دیابت تصمیم گیری کند. یک بار با استفاده از روش درخت تصمیم گیری(روش C4.5) و یک بار دیگر با استفاده از روش جنگل تصادفی معرفی شده.

داده ها را به دو دسته train_set , test_set تقسیم کنید و با استفاده از cross validation مدل نهایی را بسازید.

همچنین برای هر کدام از مدلهای ساخته شده confusion matrix را محاسبه نموده و سپس منحنی ROC را رسم کنید.

کد خود به همراه فایل توضیحات را در سامانه آموزش ایلرن بارگزاری کنید. حتما در توضیحات خود بررسی کنید کدام یک از دو مدل عملکرد بهتری در طبقه بندی داشته اند(میتوانید از منحنی ROC های رسم شده و مقایسه سطح زیر آنها استفاده کنید.)

توضیحات مربوط به مجموعه داده:

داده ها مربوط به سوابق پزشکی عده ای از سرخپوستان آمریکا است و ویژگی هایی مانند تعداد دفعات بارداری – شاخص توده بدنی – فشار خون – سن فرد و ... وجود دارد. همچنین در ستون آخر ابتلا یا عدم ابتلای فرد به بیماری دیابت بررسی شده است.(عملا دو کلاس افراد دیابتی و افراد غیر دیابتی داریم)

لینک های کمکی :

Test-Train split + cross validation

Confusion matrix (intro)

ROC curve (intro)

ROC curve (python)

Tree algorithm: C4.5

Data (description + download)