



پیش پردازش

هر یک از سوالات زیر را بر روی مجموعه داده Adults اجرا نمایید و تغییرات هر مرحله را با استفاده از دیتافریم نمایش دهید.

۱. ویژگی‌های عددی این مجموعه داده را استخراج نمایید.
۲. در کل مجموعه داده مقادیر خالی ویژگی‌های اسمی را با nan جایگزین نمایید.
۳. در مجموعه داده فوق مقادیر عددی فاقد مقدار را با مقدار میانگین آن ویژگی جایگزین کنید.
۴. با استفاده از یک روش نرمالسازی، مقادیر عددی را نرمال نمایید.
۵. مقادیر عددی ویژگی سن را به ۱۰ bins با عمق یکسان تقسیم کنید.
۶. پیوستگی و correlation بین ویژگی‌ها را بدست آورید و ویژگی‌های با کوارینانس بالا را حذف نمایید و بگویید در انتها چند ویژگی باقی خواهد ماند. این کار را یکبار برای ویژگی‌های عددی و یکبار برای کل ویژگی‌ها انجام دهید و در مورد نحوه تصمیم‌گیری خود برای حذف ویژگی‌ها توضیحات لازم را بنویسید.

مهلت تحویل: ۱۴۰۲/۰۲/۲۳

پاسخ تمرین را در قالب یک فایل PDF بصورت یک گزارش فنی (Technical Report) آماده سازید.

می‌توانید از زبان پایتون برای رسم نمودارها و پاسخ به سوالات استفاده نمایید.

اگر تمرین شامل برنامه‌نویسی است ارائه سورس کد برنامه در فایل پیوست الزامی است.

تمامی فایل‌های ارسالی به فرمت rar و نام فایل‌ها به فرمت زیر باشد که # شماره تمرین شماست:

DM_HW#_YourStudentNumber_Family.rar