به نام او..

تمرین اول درس داده کاوی

گزارش:

گزارش بایستی در قالب فایل PDF باشد. لطفا فایل Word نفرستید. در گزارش تحلیل خود را در رابطه با تمام کارهایی که انجام داده اید بیان نمایید. بخش اصلی این تمرین گزارش آن می‌باشد بنابراین عدم وجود آن معادل نمره‌ی صفر خواهد بود.

فایل گزارش خود را به شکل « Report1_StdNum.pdf» نامگذاری کنید. (مانند Report1_9131081.pdf)

کد:

کد اجرایی می تواند در محیط R یا Python تهیه شود. کد و گزارش مربوط در پوشه‌ای به نام R یا Python قرار داده شود. در صورت انجام تمرین در هر دو محیط، پوشه‌ای به نام other برای محیط دوم ساخته شود. نمره‌ی اصلی برای محیط اول و نمره‌ی مثبت برای محیط دوم در نظر گرفته خواهد شد. عدم وجود کد معادل نمره‌ی صفر خواهد بود.

فایل کد خود را به شکل « DM1_StdNum » نامگذاری کنید.

بارگذاری:

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

فایل فشرده را به شکل «DM1_StdNum» نامگذاری کنید. (مانند DM1_9131081)

مهلت ارسال تمرین ساعت ۲۳:۵۵ دقیقه‌ی روز جمعه مورخ ۶ اسفند 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.

شرح تمرین:

هدف از این تمرین آشنایی با یکی از محیط‌های R یا Python در قالب حل یک مسئله‌ی داده کاوی می باشد. برای این منظور یکی از مسابقه های آموزشی سایت Kaggle مورد استفاده قرار می‌گیرد. از این لینک می‌توانید به صفحه‌ی این مسابقه دسترسی داشته باشید. در قسمت getting-started از بخش Information لینک‌های آموزشی مختلفی برای شروع به کار قرار داده شده است. دو منبع مناسب برای هر محیط در ادامه آورده شده است. می توانید از منابع دیگر نیز استفاده کنید.

R:

1- https://www.kaggle.com/mrisdal/titanic/exploring-survival-on-the-titanic
2- http://trevorstephens.com/kaggle-titanic-tutorial/getting-started-with-r/

Python:

1- https://www.kaggle.com/omarelgabry/titanic/a-journey-through-titanic
2- http://ahmedbesbes.com/how-to-score-08134-in-titanic-kaggle-challenge.html

هدف از این مسئله آموزش مدلی برای پیش بینی زنده ماندن یا کشته شدن مسافران کشتی تایتانیک می‌باشد. انتظار می‌رود با کمک لینک های اشاره شده کد مورد نیاز برای پیش بینی نتایج بر روی داده‌ی تست و ساخت خروجی طبق فرمت مورد نیاز برای بارگذاری نتایج در سایت Kaggle را پیاده سازی نمایید. از کامنت های مناسب برای بیان بخش‌های مختلف کدتان استفاده کنید. نام کاربری و نتیجه‌ی بهترین مدل خود در سایت Kaggle را در ابتدای گزارش بیان کنید. در فایلی که در سایت درس بارگذاری می‌کنید، کد بهترین نتیجه ی خود را قرار دهید. البته بایستی توضیحی در مورد تمام مدل ها و نتایج گرفته شده در گزارش خود آورده باشید.

به پرسش‌های زیر نیز که در حین آموزش به آن‌ها برمی‌خورید پاسخ دهید:

۱) DataFrame چیست؟
۲) مقصود از Imputation چیست؟
۳) Normalize کردن داده‌ها به چه منظوری صورت می‌گیرد؟
۴) آیا در درخت تصمیم، تمام ویژگی‌های مسئله دارای اهمیت یکسانی هستند؟
۵) در الگوریتم جنگل تصادفی دو پارامتر max_depth و min_sample_split بیانگر چه چیزی هستند؟

موفق باشید