به نام او ...

تمرین پنجم درس داده کاوی

تمرکز این تمرین بر روی مبحث دسته‌بندی می‌باشد. دو مجموعه داده برای دسته‌بندی در نظر گرفته شده است. هر دو بخش این تمرین به صورت رقابتی بوده و عملکرد بهتر مدل بر روی مجموعه داده‌ی آزمایش معادل با نمره‌ی بیش‌تر خواهد بود.

بخش اول (۷۵ نمره)

مجموعه داده‌ی این بخش از تراکنش‌های آنلاین کارت‌های اعتباری تشکیل شده است. تعدادی از این تراکنش‌ها تقلب(Fraud) هستند. هدف این بخش آموزش مدلی برای پیش بینی این تقلب‌ها برای تراکنش‌های جدید می‌باشد. با مدل آموزش دیده برچسب داده‌های آزمایش را پیش‌بینی کرده و همانند برچسب داده‌های آموزش ذخیره کنید. این پیش‌بینی‌ها به دلیل نامتعادل بودن برچسب‌های مجموعه داده با دو معیار F1-Score و Accuracy بر روی قسمت متعادلی از داده‌های آزمایش مقایسه خواهند شد.

انتظار می‌رود موارد زیر در پیاده‌سازی مدل مد نظر قرار بگیرد.

۱- ویژگی‌های مجموعه داده را به طور دقیق بررسی کرده و یافته‌ها را در گزارش خود بیان کنید. برای مثال با کمی جست و جو می‌توان فهمید که ویژگی zip همان zipcode بوده و عددی یا غیرعددی بودن آن را تشخیص داد. بررسی correlation داده‌ها می‌تواند صورت بگیرد. امکان یکسان بودن دو ویژگی نیز وجود دارد که باید اقدام به حذف ویژگی تکراری نمود. داده‌ها را در ابتدا به خوبی بررسی کنید تا دید خوبی نسبت به ویژگی‌ها پیدا کنید.

۲- در این مسئله تعداد داده‌های با برچسب ۱ (Fraud) بسیار کم تر از تعداد داده‌های با برچسب ۰ می‌باشد که منجر به نامتعادل شدن داده‌ها و بایاس شدن مدل به سمت برچسب ۰ می‌شود. راه حلی برای این مشکل ارائه دهید. در صورتی که این مشکل را حل نکرده و مدلتان برای هر داده‌ای برچسب ۰ پیش بینی کند، F1-Score و Accuracy بر روی زیرمجموعه‌ی متعادل از داده‌ی آزمایش بسیار پایین خواهد بود.

۳- حداقل چهار مدل Decision Tree، Random Forest، Neural Network و Logistic Regression را با پارامترهای مختلف برای دسته بندی استفاده کرده و نتایج آن‌ها را گزارش کنید. برای مقایسه‌ی مدل‌ها از هر روش و معیار دلخواهی می‌توانید استفاده کنید. البته باید در گزارش خود دلیل استفاده و توضیح مختصری در مورد آن بیان کنید.

۴- مدل‌های مختلف را با استفاده از روش‌های موجود ترکیب کرده و عملکرد مدل را بررسی کنید.

۵- با استفاده از ویژگی‌های اصلی ویژگی‌های جدیدی ساخته و بهبود یا عدم بهبود عملکرد مدل را مورد بررسی قرار دهید.

بخش دوم (۲۵ نمره)

در تمرین اول مسئله‌ی پیش‌بینی زنده ماندن در تایتانیک صرفا برای آموزش ابزارهای R و Python مورد استفاده قرار گرفت. در این تمرین هدف بررسی دقیق تر این مسئله می‌باشد. با بهره گیری از مفاهیمی که در طول ترم در کلاس یاد گرفته‌اید، اقدام به بررسی مجدد این مسئله نمایید. در صورت بهبود عملکرد مدل، توضیحی در مورد آن در گزارش خود بیان نمایید. کسانی که تمرین ۱ را تحویل نداده‌اند با انجام دادن این بخش، قسمتی از نمره‌ی آن تمرین را نیز خواهند گرفت.

انتظار می‌رود موارد زیر را در حل این مسئله مد نظر قرار دهید.

۱- استفاده از حداقل ۳ مدل مختلف برای دسته‌بندی.
۲- ترکیب مدل‌های مختلف(Ensemble).
۳- ایجاد ویژگی‌های جدید از مجموعه داده و بررسی بهبود یا عدم بهبود عملکرد مدل.

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 مهمی از کار مربوط به فرآیند رسیدن به نتیجه ی نهایی می باشد. بنابراین گزارش بخش قابل توجهی از نمره را به خود اختصاص می دهد.

فایل گزارش خود را به شکل «Report5_StdNum.pdf» نامگذاری کنید. (مانند Report5_9131081.pdf)

کد:

کد اجرایی می تواند در محیط R یا Python تهیه شود. عدم وجود کد معادل نمره‌ی صفر خواهد بود.

فایل کد خود را به شکل «DM5_P1_StdNum» برای بخش اول و «DM5_P2_StdNum» برای بخش دوم نامگذاری کنید.

بارگذاری:

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

فایل فشرده را به شکل «DM5_StdNum» نامگذاری کنید. (مانند DM5_9131081)

مهلت ارسال تمرین ساعت ۲۳:۵۵ دقیقه‌ی روز جمعه مورخ ۱۲ خرداد 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