تمرین سری دوم درس داده کاوی

مهلت تحویل: ۱۴ فروردین ۱۴۰۱

۱. قسمت اول

- **دادگان قسمت اول:**

در این قسمت دادگان مربوط به وضعیت سلامت مادران باردار در دسترس است. این دادگان شامل ۱۰۱۴ نمونه با ۷ ستون است(ستون آخر مربوط به Label است و ۶ ستون اول مربوط به ویژگی های ورودی). میخواهیم عمل دسته بندی را روی این داده ها و در سه گروه low risk ، mid risk و high risk انجام دهیم.

از این لینک میتوانید دادگان (dataset) را دانلود کنید.

- **شرح قسمت اول:**
  - ابتدا دادگان را load کنید. Label هر داده که مربوط با ستون RiskLevel است را از ویژگی های آموزشی جدا کنید. y بردار متناظر با ستون RiskLevel و X متناظر با داده های موجود در ۶ ستون اول باشد. (باید y را از X جدا کنید) .
  - روی بردار y بدست آمده عمل Encoding را انجام دهید تا داده های categorical به عدد صحیح تبدیل شود. (مثلا به جای low risk مقدار ۰ ، به جای mid risk مقدار ۱ و به جای high risk مقدار ۲ را قرار دهید.)
  - داده ها را shuffle کنید و با نسبت ۷۰ به ۳۰ به داده های آموزشی و تست تقسیم کنید. داده های آموزشی را X_train و y_train و داده های تست X_test و y_test نامگذاری کنید.
  - پیش پردازش ها را در صورت لزوم روی داده ها انجام دهید.
  - دسته بندی را با استفاده از مدل های Decision Tree، KNN، SVM و Random Forest انجام دهید. بهترین پارامتر ها برای این مدل ها با استفاده از Cross Validation بدست آورید و آن ها را ذکر کنید. دقت کنید برای انجام Cross Validation نباید از داده های تست استفاده کنید. داده های تست نباید در فرآیند آموزش استفاده شوند. دقت ها را برای داده های آموزشی و تست و همچنین Cross Validation را در جدولی ذکر کنید. Confusion Matrix را در هر قسمت ارائه کنید. تحلیل های خود را برای نتایج بدست آمده ارائه کنید.
  - در پایان عملکرد همه ی قسمت ها را در کنار هم قرار دهید و نتیجه گیری کنید. ارائه نمودار امتیاز مثبت خواهد داشت.

۲. قسمت دوم

- **دادگان قسمت دوم:**

در این قسمت از تمرین دادگان مربوط به بیماری پارکینسون از روی صدای افراد در دسترس است. این دادگان شامل ۷۶۵ نمونه (که از ۱۸۸ فرد بیمار و ۶۴ فرد سالم و از هر کدام ۳ آزمایش گرفته شده است که جمعا برابر ۷۶۵ آزمایش است) با ۷۵۴ ستون است (۷۵۳ ستون مربوط به ویژگی و یک ستون مربوط به label است). میخواهیم عمل دسته بندی را روی این داده ها و در دو گروه ۱ و ۰ انجام دهیم.

از این لینک میتوانید دادگان (dataset) را دانلود کنید.

- **شرح قسمت دوم:**
  - ابتدا دادگان را load کنید. Label هر داده که مربوط با ستون class است را از فیچر های آموزشی جدا کنید. y بردار متناظر با ستون class و X متناظر با داده های موجود در ۷۵۳ ستون اول باشد. (باید y را از X جدا کنید) .
  - داده ها را shuffle کنید و با نسبت ۷۰ به ۳۰ به داده های آموزشی و تست تقسیم کنید. داده های آموزشی را X_train و y_train و داده های تست X_test و y_test نامگذاری کنید.
  - پیش پردازش ها را در صورت لزوم روی داده ها انجام دهید (راهنمایی: شاید حذف ستون id دقت را افزایش دهد.)
  - ابتدا بررسی کنید اگر مدلی داشته باشیم که به ازای هر ورودی خروجی ۱ بدهد دقت ما برای داده های آموزشی و تست چقدر خواهد بود. نتیجه را تحلیل کنید.
  - دسته بندی را با استفاده از مدل های Decision Tree، KNN، SVM و Random Forest انجام دهید. بهترین پارامتر ها برای این مدل ها با استفاده از Cross Validation بدست آورید و آن ها را ذکر کنید. دقت کنید برای انجام Cross Validation نباید از داده های تست استفاده کنید. داده های تست نباید در فرآیند آموزش استفاده شوند. دقت ها را برای داده های آموزشی و تست و همچنین Cross Validation را در جدولی ذکر کنید. Confusion Matrix را در هر قسمت ارائه کنید. تحلیل های خود را برای نتایج بدست آمده ارائه کنید.
  - سپس با استفاده از PCA، عمل Feature Extraction را انجام دهید و مرحله قبل را تکرار کنید. تجربیات خود را درباره تاثیر تعداد مولفه های اصلی بر روی عمل دسته بندی را در جدولی ارائه کنید و آن را تحلیل کنید. بهترین تعداد برای مولفه های اصلی را در گزارش ذکر کنید.
  - در پایان عملکرد همه ی قسمت ها را در کنار هم قرار دهید و نتیجه گیری کنید. ارائه نمودار امتیاز مثبت خواهد داشت.

- **نحوه ارسال فایل های تمرین :**
  - کد های خود را به همراه فایل مربوط به توضیحات به صورت یک فایل فشرده در سامانه ایلرن ثبت کنید.
  - ترجیحا فایل کد ارسالی به فرمت ipynb. باشد.
  - نام فایلی که آپلود می کنید شامل نام نام خانوادگی و شماره دانشجویی باشد . مثال :

DM - HW2 - your name - your student id

- **پیشنهادات و نکات مربوط به انجام تمرین :**
  - برای انجام قسمت مربوط به نوشتن کد سعی کنید از jupyter notebook استفاده کنید.
  - استفاده از پکیج ها و کتابخانه های آماده بلامانع است.
  - برای خواندن داده ها میتوانید از دستور pandas.read_csv() استفاده نمایید.

موفق باشید

اسفند ۱۴۰۰