توجه:

دقت کنید که علاوه بر فایل‌های کد برای هر سوال (هر کدام با شماره سوال مشخص شود)، باید گزارش کار نیز نوشته شود. اگر گزارش کار شما ناقص باشد پروژه شما مورد قبول نیست. فایل کد بدون گزارش کار هیچ نمره‌ای ندارد. لطفا در گزارش کار همه جزئیات و نتایج را بنویسید(در گزارش کار دوباره کدهای خود را ننویسید).

۱. دیتاست این سوال با اسم Datapreprocessing در فایل این تمرین ضمیمه شده است. قصد ما در این سوال یادگیری پیش پردازش داده‌ها است.

توضیحی در مورد دیتاست: این دیتاست در مورد چند کشور مختلف است که در آن ویژگی هایی مانند جمعیت این کشورها، رشد جمعیت و وضعیت توریستی وجود دارد که به کمک این ویژگی ها قرار است مدل رگرسیونی بسازیم که تعداد بیماران مبتلا به ویروس کرونا در کشورهای دیگر را تخمین بزنیم. البته دقت کنید با توجه به تعداد کم داده‌ها و اینکه ویژگی ها لزوما ویژگی‌های دقیقی نیست، انتظار نتیجه دقیقی نداریم و فقط و فقط هدف یادگیری است.

(آ) راه‌های مختلف مقابله با Missing Values را به کار ببرید و به نظر شما کدام یک از راه ها مناسب‌تر است؟ آیا می‌توان نظر کلی داد؟ (تمام راه‌ها در ویدیوی پیش پردازش داده‌ها و همچنین در کلاس درس گفته شده است که می‌توانید از آن‌ها استفاده کنید).

(ب) چرا نمی‌توان از Categorical Variables برای داده‌کاوی استفاده نمود و باید آن را به ویژگی‌های عددی تبدیل نمود؟ چگونه آن‌ها را می‌توان به متغیرهای عددی تبدیل نمود؟ این روش را بر روی این دیتاست اعمال کنید.

(ج) به نظر شما در این دیتاست آیا به Feature Scaling نیاز داریم یا خیر؟ چرا؟ اگر پاسخ شما مثبت است، روش‌های مختلف Feature Scaling را بر روی این دیتاست اعمال کنید و از نظر شما کدام مناسب‌تر است؟

(د) داده‌های پرت را در این دیتاست مشخص کنید و آن‌ها را حذف کنید. آیا از نظر شما حذف کردن داده‌های پرت روش درستی است؟ اگر جواب شما منفی است راه حل جایگزین ارائه کنید.

(ه) فرض کنید همه داده‌های ما داده آموزش است. Multiple Linear Regression را بر روی این دیتاست پیش پردازش شده اعمال کنید. مدل را به طور کامل گزارش کنید. منظور از گزارش مدل، مشخص کردن ضرایب، عرض از مبدا، خطاهای MSE ، RMSE و خروجی به کمک Statsmodels می‌باشد. گزارش خود را تحلیل کنید.

(و) فرض کنید ستون Total population را در متغیر X قرار دهیم و ستون تعداد مبتلایان به ویروس کرونا را در متغیر y قرار دهیم. حال می‌خواهیم مدل رگرسیونی به فرم $y = aX^2 + bX + c$ بسازیم. پارامترهای مجهول را بدست آورید.

۲. دیتاست این سوال با اسم Ecommerce Customers در فایل این تمرین ضمیمه شده است. به توضیحات زیر در مورد هدف ما در این سوال و همچنین دیتاست توجه کنید.

Congratulations! You just got some contract work with an Ecommerce company based in New York City that sells clothing online but they also have in-store style and clothing advice sessions. Customers come in to the store, have sessions/meetings with a personal stylist, then they can go home and order either on a mobile app or website for the clothes they want.

The company is trying to decide whether to focus their efforts on their mobile app experience or their website. They've hired you on contract to help them figure it out!

We'll work with the Ecommerce Customers csv file from the company. It has Customer info, suchas Email, Address, and their color Avatar. Then it also has numerical value columns:

Avg. Session Length: Average session of in-store style advice sessions.

Time on App: Average time spent on App in minutes

Time on Website: Average time spent on Website in minutes

Length of Membership: How many years the customer has been a member.

(آ) ابتدا به کمک pairplot و heatmap در مورد همبستگی ستون‌ها نظر دهید.

(ب) Multiple Linear Regression را بر روی این دیتاست اعمال کنید. مدل را به طور کامل گزارش کنید. منظور از گزارش مدل، مشخص کردن ضرایب، عرض از مبدا، خطاهای MSE داده‌ی آموزش ، RMSE داده‌ی آموزش و خروجی به کمک Statsmodels می‌باشد. گزارش خود را تحلیل کنید.

(ج) به کمک $k-$ fold cross validation خطای تست را تقریب بزنید.

(د) حال با استفاده از مدل به دست آمده، داده‌های تست را پیش بینی کنید و RMSE تست را بدست آورید. با توجه به خطای داده‌ی تست و آموزش آیا overfit یا underfit رخ داده است. نظرتان را در مورد bias و variance این مدل بیان کنید. اگر مدل به خوبی کار نکرده است، دلایل آن را بیان کنید. (البته دقت کنید در مسائل واقعی ما هیچ وقت به مقدار متغیر هدف دیتای تست دسترسی نداریم و باید با تقریب خطای تست مدل خود را ارزیابی کنیم)

(ه) حال با توجه به ارزیابی خود از مدل به جوابی که شرکت از شما دارد پاسخ دهید.

Do you think the company should focus more on their mobile app or on their website?

۳. در این قسمت می‌بایست در خروجی همه سوالات معیارهای ROC، Confusion Matrix و Classification Report بیان شود.

(آ) دیتاست

http://archive.ics.uci.edu/ml/datasets/Arrhythmia

را دانلود کنید. همان طور که مشاهده می‌شود این دیتاست دارای Missing Values می‌باشد. با یکی از راه‌های مقابله با داده گمشده، پیش پردازش داده را انجام دهید.

(ب) الگوریتم نزدیک ترین همسایه را برای $k = ۱$ و $k = ۳۰$ با معیار فاصله اقلیدوسی، بر روی دیتاست دانلود شده اعمال کنید. تمام معیارهای آموزش را برای این داده‌های آموزش و تست بدست آورید و آن‌ها را تحلیل کنید.

(ج) به کمک روش $\mathrm{k-fold\ cross\ validation}$ مقدار بهینه k و بهترین معیار فاصله (کسینوسی،اقلیدوسی و منهتن) را بدست آورید. حال این الگوریتم را به ازای این مقادیر بهینه بر روی دیتاست دانلود شده اعمال کنید. تمام معیارهای آموزش را برای داده‌های آموزش و تست بدست آورید و آن‌ها را تحلیل کنید.