پروژه دوم: یادگیری ماشین مهلت تحویل: ۲۴دی‌ماه ۱۴۰۰

این پروژه دارای دو بخش است و برای هر بخش یک دیتاست جداگانه در نظر گرفته شده است و هر گروه برای این پروژه باید موارد خواسته شده را بر روی داده‌ها پیاده‌سازی کند.

بخش اول: طبقه‌بندی

در این بخش شما باید بر روی یک دیتاست مربوط به بیماران قلبی یک کلینیک پزشکی در کلیولند آمریکا کار کنید. داده‌های اصلی دارای ۷۶ مشخصه بوده‌اند اما در همهٔ آزمایشات انجام شده فقط ۱۴ مشخصه کاربردی شناخته شده‌اند که این ۱۴ مشخصه در فایل Heart.csv پیوست پروژه قابل‌مشاهده است.

مشخصه‌های موجود در دیتاست (با توجه به هدف این پروژه و موارد خواسته شده، برای انجام پروژه نیازی به فهم دقیق این مشخصه‌ها و معنای آن‌ها نیست):

۱) age:سن بیمار
۲) sex:جنسیت بیمار
۳) cp: chest pain type (four values)
   -- Value ۱: typical angina
   -- Value ۲: atypical angina
   -- Value ۳: non-anginal pain
   -- Value ٤: asymptomatic
۴) trestbps: resting blood pressure (in mm Hg on admission to the hospital)
۵) chol: serum cholesterol in mg/dl
۶) fbs: (fasting blood sugar > ۱۲۰ mg/dl) (۱ = true; ۰ = false)
۷) restecg: resting electrocardiographic results (values ۰,۱,۲)
   -- Value ۰: normal
   -- Value ۱: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > ۰,۰۵ mV)
   -- Value ۲: showing probable or definite left ventricular hypertrophy by Estes' criteria
۸) thalach: maximum heart rate achieved
۹) exang: exercise induced angina (۱ = yes; ۰ = no)
۱۰) oldpeak: ST depression induced by exercise relative to rest
۱۱) slope: the slope of the peak exercise ST segment

-- Value ۱: upsloping
-- Value ۲: flat
-- Value ۳: downsloping

۱۲) ca: number of major vessels (۰-۳) colored by flourosopy
۱۳) thal
۱۴) target: diagnosis of heart disease (angiographic disease status)
-- Value ۰: No heart disease
-- Value ۱: Have heart disease

هدف این است تا مدلی بسازید که بتواند بر اساس ۱۳ متغیر اول پیش‌بینی کند که آیا بیمار دچار بیماری قلبی هست یا خیر، برای این کار گام‌های زیر را طی کنید.

- ابتدا به کمک کتابخانه pandas داده‌ها را در نوت‌بوک بخوانید و ۱۰ ردیف اول را نمایش دهید.
- آیا در داده‌ها مقدار گمشده[۱] وجود دارد (برای فهمیدن این موضوع می‌توانید از دستور ()isna بر روی دیتا فریم خود استفاده کنید)؟ در صورت مثبت بودن پاسخ سطر مربوط به آن را از داده‌ها حذف کنید.
- به کمک کتابخانه seaborn نمودار heatmap را برای داده‌ها رسم کنید و تحلیل خود را از این نمودار بنویسید.
- ۱۳ ستون اول را به‌عنوان متغیر پیش‌بینی‌کننده و ستون آخر یعنی target را به‌عنوان متغیر پاسخ در نظر بگیرید.
- متغیرهای X (پیش‌بینی‌کننده) را استاندارد کنید.
- داده‌های خود را به دو قسمت test و train تقسیم کنید (می‌توانید نسبت‌های مختلفی را برای test و train امتحان کنید و تأثیر آن را در نتیجه نهایی مدل‌سازی خود بررسی کنید، مثلاً یکبار ۲۰ درصد داده‌ها را برای تست بگذارید و یکبار ۱۰ درصد و در نهایت مقداری که به‌ازای آن مدل‌سازی نتیجه بهتری می‌دهد انتخاب کنید).
- با روش K-NN و به‌ازای یک مقدار K دلخواه بر روی داده‌های train مدل‌سازی انجام دهید، به کمک مدل خود مقادیر y را برای داده‌های test پیش‌بینی کنید و دقت مدل را بر روی داده‌های train و داده‌های test محاسبه کنید و تحلیل خود را از آن بنویسید.
- آیا مقداری که برای k در نظر گرفته‌اید بهترین مقدار بوده است؟ با استفاده از یک حلقه، مدل‌سازی را به‌ازای مقادیر مختلف برای k انجام دهید و بهترین مقدار برای k را انتخاب کنید.

[۱] Missing Value

- حالا یکبار دیگر مدل‌سازی را به‌ازای مقدار بهینه k انجام دهید و ماتریس آشفتگی، precision، recall و f۱-score را برای مدل محاسبه کنید و تحلیل خود را دررابطه‌با مقادیر به‌دست‌آمده برای precision، recall و f۱-score بنویسید.
- در صورت تمایل یکی از روش‌های Logistic Regression، SVM یا Decision Tree را انتخاب کرده، توضیح مختصری درباره آن بنویسید و مدل‌سازی را بر روی داده‌ها با یکی از این روش‌ها انجام دهید و نتیجه آن را با روش K-NN مقایسه کنید. [۱] (**این گام اختیاری است و نمره اضافی برای آن در نظر گرفته می‌شود بنابراین الزامی به انجام آن نیست.)**

بخش دوم: خوشه‌بندی

در این بخش هدف تقسیم‌بندی مشتریان است. فرض کنید که شما مالک یک کسب‌وکار هستید و یک سری اطلاعات پایه راجع به مشتری‌های خود دارید. این اطلاعات در فایل Customers.csv پیوست پروژه قابل‌مشاهده است.

مشخصه‌های موجود در دیتاست:

۱) CustomerID: کد مشتری
۲) Gender:جنسیت مشتری
۳) Age: سن مشتری
۴) Annual Income (k$): درآمد مشتری به ۱۰۰۰ دلار
۵) Spending Score (۱-۱۰۰): امتیاز مشتری از ۱ تا ۱۰۰ بر اساس خریدهایی که انجام داده است

شما می‌خواهید بر اساس این داده‌ها مشتری‌های خود را به چند دسته تقسیم کنید تا بتوانید به‌صورت هدفمندتر برای هر دسته بازاریابی انجام دهید. برای پیداکردن دسته‌های مختلف و ویژگی‌های هر دسته گام‌های زیر را طی کنید:

- ابتدا به کمک کتابخانه pandas داده‌ها را در نوت‌بوک بخوانید و ۱۰ ردیف اول را نمایش دهید. ازآنجایی‌که ما در روش K-Means فاصله نقاط را محاسبه می‌کنیم نیاز است که همهٔ داده‌ها عددی باشند. متغیرهایی که به شکل عددی نیستند را به شکل عددی تغییر دهید. برای مثال در متغیر جنسیت

[۱] در صورت تمایل برای یادگیری هر یک از روش‌های ذکر شده می‌توانید از کورس رایگان زیر استفاده کنید:
https://www.coursera.org/learn/machine-learning-with-python

مرد و زن را با ۱ و ۰ جایگزین کنید. برای این کار روش‌های مختلفی وجود دارد که یکی از آن‌ها به‌صورت زیر است:

```
df ['Gender'] = df['Gender'].map({'Male': ۱, 'Female': ۰})
```

- برای شناخت بهتر داده‌ها مقادیر Max، Min، میانگین و انحراف معیار را برای متغیرها محاسبه کنید (برای این کار می‌توانید از دستور describe() بر روی دیتاست خود استفاده کنید).
- متغیرهایی که می‌خواهید بر اساس آن‌ها خوشه‌بندی را انجام دهید مشخص و استاندارد کنید. (دقت کنید که ممکن است همه متغیرها لازم نباشند، برای مثال CustomerID اطلاعات ارزشمندی را از مشتری به شما نمی‌دهد)
- روش K-Means را به‌ازای مقدار دلخواه برای K پیاده‌سازی کنید سپس یک ستون به دیتاست خود اضافه کنید که در آن، cluster به‌دست‌آمده برای هر مشتری نمایش داده شود و ۵ سطر اول دیتاست را در حالت جدید نمایش دهید.
- میانگین متغیرهای مختلف را به‌ازای هر یک از clusterها محاسبه کنید. (برای این کار از دستور groupby() استفاده کنید) تحلیل خود را برای هر یک از clusterها بنویسد برای مثال اینکه میانگین سنی، درآمدی و ... در هر cluster به چه شکل است.
- برای مشاهده و درک بهتر clusterها نمودارهایی مشابه با آنچه در کلاس حل تمرین رسم شد برای داده‌های خود رسم کنید.
- آیا مقداری که برای K در نظر گرفتید، مقدار مناسبی بوده است؟ با استفاده از قانون elbow مشخص کنید که مقدار مناسب برای k چند است؟

نکته‌های تحویل

۱) توضیحات هر گام و تحلیل‌هایی که در هر گام انجام می‌دهید و نمودارهایی که رسم می‌کنید به همراه تحلیل مختصری دررابطه‌با هر نمودار را در یک فایل ورد بنویسید. (در صورت تمایل می‌توانید توضیحات خود را در همان فایل نوت بوک و با استفاده از قابلیت Markdown بنویسید، در اینصورت نیازی به فایل ورد نیست).
در صورتی که گزارش خود را در فایل ورد می‌نویسد نیازی به آوردن کدهای نوشته شده در فایل ورد نیست. فایل مربوط به کدهای خود را به پیوست فایل گزارش خود آپلود کنید.

۲) به‌ازای هر روز تأخیر ۱۵% از نمره کسر می‌شود و نهایتاً ۳۰% نمره باقی می‌ماند.

۳) ۱۰% از نمره مربوط به انشای درست، نداشتن غلط املایی و مرتب بودن قالب گزارش است. نکته‌های نگارشی موردنظر در صفحه درس در کوئرا ارسال شده است و انتظار می‌رود تمامی موردها رعایت شوند.

۴) مشکلات و سؤالات خود را در صفحه مربوط به این پروژه در صفحه کلاس در کوئرا مطرح کنید.