دانشگاه اصفهان

تمرین شماره ۱

یادگیری ماشین در تجارت الکترونیک

موعد تحویل: **یکشنبه ۲۸ آبان ۱۴۰۲**

نام استاد: **دکتر مرجان کائدی**

دستیار حل تمرین: **رضا شیری**

پاییز ۱۴۰۲

نکات مورد نیاز درباره‌ی تمرین

- فایل‌های ارسالی شما باید شما دو بخش کدهای پیاده‌سازی شده و همچنین یک گزارش فنی شامل توضیحات مربوط به تمرین باشد.
- در صورت مشاهده‌ی مشابهت‌های غیر طبیعی، **کسر نمره** به افراد تعلق می‌گیرد.
- در صورت استفاده از یک وب‌سایت یا منبع خاص، در فایل گزارش خود آدرس وب‌سایت یا نام منبع را ذکر کنید. در صورت عدم ذکر این موارد و پیدا شدن مشابهت‌های غیر طبیعی، کسر نمره تعلق می‌گیرد.
- شما می‌توانید حداکثر به‌صورت دو نفره، تمرین را تحویل دهید.
- پاسخ‌های خود را به صورت یک فایل zip به فرمت [name1]_[family1]_[name1]_[family1] [#assignment]_ [std#2]_ [std#1]_ در آورید که در آن [name] نام، [family] نام‌خانوادگی، [assignment#] شماره‌ی تکلیف و [std#] شماره دانشجویی می‌باشد. (به‌طور مثال Reza_Shiri_953611133047_Siavosh_Djazmi_993624008_#1)
- پاسخ‌های خود را تا **ساعت ۲۳:۵۹** روز **یکشنبه ۲۸ آبان** در قسمت مربوط به تمرین ۱ کلاس تعریف شده در سامانه‌ی کوئرا ارسال کنید.
- مهلت پاسخ‌گویی به این تمرین تا پایان روز یکشنبه ۲۸ آبان می‌باشد. با توجه به اختلالات مربوط به اینترنت و مشکلات دیگری که برای شما ممکن است پیش بیاید، شما می‌توانید تا **حداکثر ۵ روز** پس از اتمام مهلت ارسال تمرین، پاسخ‌های خود را **بدون کسر نمره**، در سامانه‌ی کوئرا آپلود کنید. توجه داشته باشید که **این زمان به‌هیچ عنوان قابل تمدید نمی‌باشد**.
- همچنین شما می‌توانید سوالات احتمالی خود را از طریق ایمیل rezamdd1998@gmail.com یا آیدی تلگرامی (@creation_bug) بپرسید. (توجه داشته باشید این ایمیل، فقط برای پاسخ‌گویی به سوالات مربوط به تمرین است و ارسال تکالیف به این آدرس نمره‌ای را به همراه نخواهد داشت.

دانشکده‌ی مهندسی کامپیوتر
گروه فناوری اطلاعات

پیش‌پردازش داده‌ها (۶ سوال)

مجموعه‌داده‌ی مورد استفاده در این تمرین، که به‌همراه تمرین ضمیمه شده است، مجموعه‌داده‌ی **Customer Classification** است. این مجموعه‌داده شامل اطلاعات مربوط به مشتریان (مانند منطقه‌ی زندگی، وضعیت بازنشستگی، وضعیت تأهل، میزان درآمد و...) است. فایل Telecust1.csv مجموعه‌داده‌ی اصلی است که داده‌های آن کامل هستند، اما یک فایل دیگر با نام Telecust1.csv -Null.csv هم در کنار فایل قرار دارد که برخی از اطلاعات مربوط به مجموعه‌داده‌ی اصلی در آن NaN[1] هستند. شما در این تمرین باید با هر دو مجموعه‌داده کار کنید.

هدف از این تمرین، پیش‌بینی دسته‌ی مشتریان است که نوع دسته‌ی آن‌ها در ستون custcat قرار دارد.

توصیه می‌شود برای انجام این تمرین از زبان پایتون استفاده کنید، اما استفاده از زبان‌های برنامه‌نویسی دیگر (مانند MATLAB، R و...) بلا مانع است. در نظر داشته باشید که درصورت استفاده از زبان‌های برنامه‌نویسی دیگر، عملکرد توابعی که استفاده شده‌اند را به‌طور خیلی مختصر توضیح دهید.

در زبان پایتون، کتابخانه‌های مورد نیاز برای انجام این تمرین، numpy، pandas، sklearn و matplotlib هستند که باید قبل از استفاده، ابتدا آن‌ها را نصب کنید.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import sklearn
```

از آن‌جایی که در این تمرین، قرار است تعداد زیادی نمودار رسم شود، بهتر است یک تابع برای رسم نمودار بنویسید و از آن در طول انجام تمرین استفاده کنید. همچنین شما می‌توانید از دستور `plt.savefig(path)` نمودارهای رسم شده را ذخیره کنید تا آسان‌تر بتوانید از آن‌ها در گزارش خود استفاده کنید.

هر دو مجموعه‌داده‌ی ضمیمه‌شده را بخوانید. از دستور زیر می‌توانید برای خواندن مجموعه‌داده‌ی موردنظر استفاده کنید.

```
df = pd.read_csv(file_path)
```

سپس با استفاده از LabelEncoder در کتابخانه‌ی sklearn، داده‌های رشته‌ای را به عددی تبدیل کنید.

۱- برای برطرف کردن رکوردهای با مقدار خالی، دو راه ۱- حذف سطرهای حاوی داده‌ی NaN و ۲- جایگزینی با مقدار میانگین ستون وجود دارد. هر دو روش را انجام دهید و تفاوت تعداد داده‌ها را گزارش کنید.

برای مراحل بعدی از مجموعه‌داده‌ی حاصل از روش دوم استفاده کنید.

[1] Not a Number

۲- ضرایب همبستگی[1] را برای ستون‌های مختلف به‌دست آورید و نتایج را تحلیل کنید. رابطه‌ی سه ستون وضعیت تأهل، میزان درآمد و tenure را (دو به دو با یکدیگر) در یک نمودار رسم کنید و نمودار به‌دست آمده را با نتایج به‌دست آمده در مرحله‌ی قبل مقایسه و تحلیل کنید.

۳- دو ستون درآمد و tenure را با ۳ روش‌های مختلف (که در اسلایدهای درس به آن‌ها اشاره شده است) نرمال‌سازی کنید، نمودار آن‌ها با یکدیگر را رسم کنید و ۳ نمودار را با نمودار قبل از نرمال‌سازی مقایسه کنید.

۴- ستون سابقه‌ی کاری (Employee) را با استفاده از روش smoothing by bin means با تعداد دسته‌ی دلخواه تغییر دهید و میانگین دسته‌ها را گزارش کنید.

۵- درباره‌ی ماتریس درهم‌ریختگی[2] و معیارهای موجود در آن (صحت، دقت و...) تحقیق کنید و به‌طور خلاصه آن‌ها را گزارش دهید.

۶- با استفاده از تابع train_test_split موجود در کتابخانه‌ی sklearn، ۸۰ درصد داده‌ها را به داده‌ی آموزشی و ۲۰ درصد به داده‌ی آزمایشی اختصاص دهید. (shuffle=False)

درخت تصمیم (۴ سوال)

۷- با استفاده از کتابخانه‌ی sklearn، مدل DecisionTreeClassifier را بارگزاری کنید، سپس با استفاده از تابع fit_transform مدل را آموزش دهید و سپس ماتریس درهم‌ریختگی را گزارش کنید و تحلیل کنید. شما برای این کار می‌توانید از classification_report و confusion_matrix که در کتابخانه‌ی sklearn موجود هستند، استفاده کنید.

۸- یک‌بار دیگر مرحله‌ی ۷ را برای مجموعه‌داده‌ای که در مرحله‌ی ۱ داده‌های NaN را حذف کردید، انجام دهید و نتایج را گزارش دهید.

۹- برای مجموعه‌داده‌ی کامل (Diabetes.csv) هم مرحله‌ی ۷ را دوباره تکرار کنید و نتایج را با مرحله‌ی قبل مقایسه کنید.

۱۰- **(این مرحله اختیاری است و می‌تواند شامل نمره‌ی اضافی برای شما باشد.)** شما می‌توانید روش‌های مختلفی که در درس با آن‌ها آشنا شدید (کاهش ویژگی، استخراج ویژگی، نرمال‌سازی‌ها، گسسته‌سازی و...) یا روش‌های دیگری که خود با آن‌ها آشنایی دارید را بر روی مجموعه‌داده‌ی اولیه و یا با تغییر پارامترهای دیگر در آموزش درخت تصمیم (مانند تغییر criterion و...) تغییرات ایجاد کنید و سپس با استفاده از آن‌ها درخت تصمیم خود را آموزش دهید. پس از هر بار آموزش نتایج را گزارش کنید. گرفتن نتایج بهتر در ماتریس درهم‌ریختگی علاوه‌بر نمره‌ی این بخش، شامل نمره‌ی اضافی‌تر برای شما نیز می‌شود.

[1] Correlation coefficients
[2] Confusion matrix