به نام خدا

تکلیف دوم درس مبانی داده کاوی

ترم دوم ۱۴۰۰-۱۴۰۱

راهنمایی :
زبان برنامه نویسی سوالات پایتون است.
پیشنهاد می شود از محیط Jupyter notebook استفاده کنید.
پکیج های اصلی مورد نیاز شامل pandas، numpy می باشند.
مجموعه داده های مورد نیاز در ادامه معرفی شده اند.
روش تحویل :
a) فایلهای مربوط به کدهای هر سوال در یک فایل با نام Qx.zip که x شماره سوال است زیپ شوند، سپس کلیه این فایلهای زیپ در یک فایل واحد با نام HW1-Lastname-StudentCode.zip که Lastname نام خانوادگی و StudentCode شماره دانشجویی شما است، زیپ شده و روی سامانه تا زمان مشخص شده آپلود شوند.
ب) گزارش نهایی باید شامل پاسخ تمامی سوالات (سوالات تشریحی و سوالات پیاده سازی) باشد که برای سوالات پیاده سازی شامل کد نوشته‌شده، توضیحی در مورد کد و نتیجه اجرا و تفسیر نتیجه می‌باشد (گزارش سوالات پیاده سازی را میتوانید در همان محیط notebook jupyter بنویسید).
ج) زمان و نحوه تحویل تکلیف در فایل راهنمای ترم مشخص شده است.
د) تحویل خارج سامانه و خارج ساعت مشخص شده قابل قبول نیست.

۱. مجموعه داده "credit cards.csv" که شامل اطلاعات پرداخت مشتریان با استفاده از کارت اعتباری است در اختیار شما قرار گرفته است. این مجموعه داده شامل اطلاعات مشتری مانند سن, جنسیت و همچنین تاریخچه ٦ ماهه برای پرداختهای هر مشتری است. هدف پیش بینی عادی بودن و یا نبودن وضعیت پرداخت مشتری است.
اطلاعات بیشتر در مورد این مجموعه داده در فایل creditCardReadMe.txt ضمیمه شده است. (زمان تقریبی لازم ۱,۵ ساعت)

با توجه به این مجموعه داده به سوالات زیر پاسخ دهید:

a. با استفاده از heatmap میزان همبستگی بین هر دو ویژگی مجموعه داده را به دست آورده و نتایج را تفسیر کنید.
b. با توجه به قسمت قبل, ۶ ویژگی که بیشترین میزان همبستگی را دارند را در یک دیتافریم جدید ذخیره کرده, و این بار به کمک نمودار pairplot همبستگی بین دو به دوی این ویژگی ها (ویژگی های ذخیره شده در دیتا فریم جدید) را به دست آورده و نتایج را تفسیر کنید.
c. به کمک روش spearman, rank correlation بین هر یک از ویژگی های مجموعه داده را با ویژگی هدف (با نام default) به دست آورده, نتایج را تفسیر کنید و بگویید کدام ویژگی بیشترین همبستگی را با ویژگی هدف دارد.
d. به کمک روش kendall, rank correlation بین هر یک از ویژگی های مجموعه داده را با ویژگی هدف (با نام default) به دست آورده, نتایج را تفسیر کنید و بگویید کدام ویژگی بیشترین همبستگی را با ویژگی هدف دارد.
آیا نتیجه با حالت قبل تفاوتی دارد؟
e. با استفاده از heatmap میزان rank correlation بین هر دو ویژگی مجموعه را با روش spearman نشان داده و نتایج را با قسمت a مقایسه کنید.

. مجموعه داده ("heart diagnose.csv") در اختیار شما قرار گرفته است. این مجموعه داده شامل اطلاعات بیماران قلبی مثل سن, جنسیت و ... و نیز اطلاعات پزشکی آنها مانند نوع درد و ... است. هدف این مجموعه داده

پیش بینی بیماری قلبی با برچسب "heart disease" است. اطلاعات بیشتر در فایل heartReadMe.txt موجود است. به کمک این مجموعه داده به سوالات ۲ تا ۴ پاسخ دهید:

۲. بررسی روشهای متقارن کردن داده (۴۵ دقیقه)

a. هیستوگرام ویژگی resting blood pressure را رسم کره و مشخص کنید که آیا کجی دارد؟ از چه نوعی است؟ مقدار عددی آن را نمایش دهید.
b. به کمک لگاریتم طبیعی سعی کنید کجی را برطرف کنید. ویژگی نرمال شده را در یک متغیر جدید ذخیره کرده و هیستوگرام و مقدار عددی کجی آن را نشان دهید.
c. به کمک جذر گرفتن سعی کنید کجی را برطرف کنید. ویژگی نرمال شده را در یک متغیر جدید ذخیره کرده و هیستوگرام و مقدار عددی کجی آن را نشان دهید.
d. یک روش جدید برای برطرف کرد کجی پیدا کنید و به کمک آن کجی را برطرف کنید. ویژگی نرمال شده را در یک متغیر جدید ذخیره کرده و هیستوگرام و مقدار عددی کجی آن را نشان دهید. ۳ روش انجام شده را با هم مقایسه کنید.

۳. کشف رابطه بین متغیرها به کمک روش overlay

(۱,۵ ساعت)

a. نمودار overlay بین ویژگی pain type و لیبل داده را در حالت غیر اسکیل شده رسم کرده و بگویید در این حالت چه رابطه ای بین این دو ویژگی کشف میکنید.
b. نمودار قسمت a را این بار به صورت اسکیل شده رسم کنید و برداشت خود را تحلیل کرده و با قسمت قبل مقایسه کنید.
c. این بار برای کشف رابطه بین این دو ویژگی از روشcomparative piechart استفاده کرده, درصدها را نشان داده و نتایج و روابط کشف شده را تحلیل کنید.
d. به کمک cross table استاندارد بین دو ویژگی بالا روابط را کشف کنید.
e. به کمک contingency table استاندارد بین دو ویژگی بالا تحلیل را انجام دهید.
f. مراحل a تا e را بین ویژگی thal و لیبل داده تکرار کنید و نتایج را تحلیل و روابط را کشف کنید.
g. یک scatter plot رسم کنید که ستون افقی آن resting blood pressure و ستون عمودی آن cholestoral باشد و پراکندگی ویژگی هدف (heart disease) را در آن نمایش دهید. آیا بر اساس این نمودار رابطه خاصی بین هر یک از این دو ویژگی با ویژگی هدف کشف میکنید؟ تحلیل خود را از خروجی نمودار بیان کنید.

۴. سبد بندی با نگاهی بر متغیر خروجی (۱,۵ ساعت)

a. نمودار overlay بین ویژگی عددی سن و لیبل داده را در حالت غیر اسکیل شده رسم کرده و بگویید در این حالت چه رابطه ای بین این دو ویژگی کشف میکنید.
b. نمودار قسمت a را این بار به صورت اسکیل شده رسم کنید و برداشت خود را تحلیل کرده و با قسمت قبل مقایسه کنید.
c. با توجه به برداشت خود از نمودار قسمت ب و کشف رابطه بین سن و بیماری قلبی و مرزهای به دست آمده, داده سن را به کمک تابع cut در پایتون سبد بندی(binning) کنید. و در یک ستون جدید با نام "age_binned" به مجموعه داده اضافه کنید.
d. نمودار overlay بین ویژگی سن سبد بندی شده و لیبل داده را در حالت غیر اسکیل شده رسم کرده و برداشت خود را از رابطه این دو بیان کنید.
e. نمودار قسمت قبل را این بار به صورت اسکیل شده رسم کنید و برداشت خود را تحلیل کرده و با قسمت قبل مقایسه کنید.

۵. با استفاده از مجموعه داده Kaggle titanic به سوالات زیر پاسخ دهید: (۶ ساعت)

- پیش پردازش
  a. تعداد missing برای هر ستون را مشخص کنید.
  b. مقادیر نامشخص ستونهای Age , Embarked و Cabin را به ترتیب با مقادیر median , mode و حذف کردن برطرف کنید.
  c. ستون جدید به نام FamilySize ایجاد کنید که برابر مجموع SibSp و Parch و خود فرد باشد.
  d. ستون Age را برای مقادیر کمتر از ۱۰ برابر با child و برای سایر مقادیر female وmale قرار دهید.
  e. ستونهای PassengerId, Name, Ticket, SibSp, Parch از مجموعه داده حذف کنید.

- **Mutual Information**
  f. بدون استفاده از کتابخانه تابع محاسبه mutual information بین مقادیر x و y را پیاده سازی کنید.
  g. بیشترین و کمترین خصوصیت مرتبط با خصوصیت Survived را مشخص کنید.
- آماده سازی برای مدلسازی
  h. خصوصیاتی با نوع categorical را به روش oneHot Encoding به عدد تبدیل کنید.
  i. همه ستونها بجز Survived را x و Survived را y تعریف کنید.
  j. داده ها را به دوبخش train و test تفکیک کنید. ۷۰٪ برای train و ۳۰٪ برای test اختصاص دهید. تا انتها تغییری در مجموعه های train و test بدست آمده ندهید. (اندازه test کمتر از ۰,۳ داده ها نباشد).
- **Decision Tree**
  k. مدل درخت تصمیم را با random_state=۰ روی داده اعمال کنید و دقت بدست آمده برای train و test را به تفکیک مشخص کنید.
  l. درخت بدست آمده را با graphViz و pydotplus نمایش دهید.
  m. آیا overfitting رخ داده است؟ چرا ؟ اگر جواب شما مثبت است راه حل این مشکل چیست؟
  n. ماتریس Confusion Matrix را برای test رسم کنید و آنرا تفسیر کنید.
- **Hyper Parameters Tunning**
  o. با استفاده از Grid Search بهترین مقدار برای پارامترهای max_depth و min_samples_leaf را بدست آورید. آیا دقت مدل بهتر خواهد شد ؟ توضیح دهید.
  p. بهترین درخت بدست آمده توسط Grid Search را با استفاده از GraphViz نمایش دهید.

- **Random Forest**
  q. مدل Random Forest را بر روی داده های train اجرا کنید.
  r. دقت بدست آمده بر روی train و test نسبت به درخت تصمیم چقدر تغییر کرده است؟
  s. با استفاده از Random Search بهترین حالت برای پارامترهای , min_samples_split , max_depth , min_samples_leaf, bootstrap و n_estimators و criterion را بدست آورید؟
  t. آیا معیار gini موثر است یا entropy ؟

*** چنانچه بتوانید برای داده test به دقت بیشتر از ۸۵٪ برسید نمره اضافه خواهید گرفت.