دانشگاه صنعتی اصفهان
دانشکده مهندسی برق و کامپیوتر

داده کاوی

تمرین سری ۳

بهار ۱۴۰۳

فهرست مطالب

۱ سوالات

۱.۱ سوال ۱

برای موارد زیر، مشخص کنید چه نوعی از یادگیری استفاده می شود ؟ (با نظارت یا بی نظارت)

(آ) پیشنهاد خرید کالاهای مشابه در سایت فروشگاه های اینترنتی

(ب) پیش بینی قیمت سهام

(ج) انتخاب مشتریان برای اعطای وام توسط بانک توسط داده های جمعیتی و مالی

(د) سیستم تشخیص پلاک

(ه) شناسایی ایمیل های هرزنامه

(و) تشخیص ناهنجاری ترافیک در شبکه جهت افزایش امنیت

(ز) تشخیص فرار مالیاتی و تقلب

(ح) نمایش پست های مورد علاقه کاربر در Explore اینستاگرام

(ط) شناسایی نژاد یک گیاه

(ی) تخمین زمان تعمیر مورد نیاز هواپیما بر اساس نام مشکل به وجود آمده

۲.۱ سوال ۲

با ذکر مثال به طور واضح توضیح دهید که Gini ratio(Gini index) و Gain ratio چه تفاوتی باهم دارند و هرکدام در کجا استفاده می‌شوند؟

۳.۱ سوال ۳

در جدول داده های سفر ۱۰ مسافر را جمع آوری کرده‌ایم. درصورتی که ستون شهر، هدف ما باشد؛ با استفاده از Gini Index درخت تصمیم را ایجاد کنید. در نظر داشته باشید برای رسم درخت تصمیم لازم است تا ابتدا بهترین ویژگی را به ترتیب با به دست آوردن Gini Index مشخص کنید.

فصل	امکان مرخصی	خودروی شخصی	شهر
تابستان	بله	بله	شیراز
تابستان	نه	بله	تهران
بهار	بله	بله	شیراز
پاییز	بله	نه	شیراز
پاییز	نه	بله	اصفهان
پاییز	بله	نه	شیراز
بهار	نه	نه	شیراز
بهار	نه	بله	رشت
بهار	بله	بله	شیراز
تابستان	نه	بله	تهران

۴.۱ سوال ۴

توضیح دهید چرا هرس کردن در استنتاج درخت تصمیم گامی سودمند است؟ ایراد استفاده از یک مجموعه مجزا از تاپل ها برای ارزیابی هرس در چیست؟

۵.۱ سوال ۵

در فایل Social_Network_Ads.csv داده های تبلیغات در شبکه های اجتماعی را در اختیار داریم. ستون Purchased در این دیتاست متغییر هدف ما می باشد. مراحل زیر را به ترتیب انجام دهید.

۱. دیتاست را با استفاده از کتابخانه pandas بخوانید و درون یک متغیر مثلا با نام dataframe قرار دهید.

۲. با استفاده از کد زیر ستون جنسیت را به نوع صفر و یک تغییر دهید.(بحث مربوط به چرایی این بخش در فصل پیش پردازش داده ها به طور مفصل توضیح داده خواهد شد.)

```
from sklearn.preprocessing import LabelEncoder
label_encoder = LabelEncoder()
dataframe['Gender'] = label_encoder.fit_transform(dataframe['Gender'])
```

۳. حالا می خواهیم متغییر های پیشگو و وابسته را مشخص کنیم. در اینجا همان طور که گفتیم متغییر Purchased وابسته و هدف است.

سایر ستون ها را در این dataframe یعنی Gender Age EstimatedSalary را با تابع iloc درون متغیر جدیدی به نام X قرار دهید. ستون Purchased را هم در متغیری به نام y قرار دهید.

با print کردن متغیر ها از درستی انجام مراحل اطمینان حاصل کنید.

۴. ضمن فراخوانی کتابخانه زیر، داده های تست و آموزشی را با درصد ۷۰ برای داده های آموزشی، افراز کنید. [1]

[1] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html

```
from sklearn.model_selection import train_test_split
```

۵. با مطالعه این لینک [۲] یک مدل درخت تصمیم ایجاد کنید و سپس آن را با داده های آموزشی که در بخش قبلی افراز کردید، fit کنید.

۶. سپس پاسخ برای مجموعه داده آزمایشی را پیش بینی (predict) کنید. [۳]

۷. حالا با استفاده از metrics [۴] دقت مدل را اندازه گیری کنید.

۸. از قطعه کد زیر برای مصورسازی درختی که ایجاد کرده اید کمک بگیرید.

```
from matplotlib import pyplot as plt
from sklearn import tree

fig = plt.figure(figsize=(25,20))
_ = tree.plot_tree(clf, filled=True)
```

توجه کنید که clf نام متغیری است که مدل شما در آن ذخیره شده است. بهتر است از دیگر ویژگی های کتابخانه matplotlib برای مصورسازی بهتر درخت خود استفاده کنید. از نمونه کد هایی که برای این کار در ژورنال های مختلف موجود است استفاده کنید.

۹. قطعه کدی بنویسید که با دریافت مقادیر Gender Age EstimatedSalary با استفاده از مدلی که ایجاد کرده اید مقدار Purchased را پیش بینی کند.

از متُد predict که در بخش های قبلی معرفی شد برای این کار استفاده کنید.

۱۰. به بخش ۵ بر می گردیم. در DecisionTreeClassifier که مدل را با آن ایجاد کردید Parameter های مختلفی از جمله criterion ، splitter ، max_depth و ... قرار دارد.

با مطالعهٔ اسناد این کتابخانه و مطالبی که در درس آموختید، توضیح دهید هر کدام از این موارد چه تاثیری بر مدل شما دارد. [۵] [۶]

۱۱. حالا با تغییر Parameter های مختلفی که در بخش قبلی معرفی شد سعی کنید دقت مدل خود را افزایش دهید. به فردی که بیشترین دقت را ارائه دهد نمره امتیازی تعلق می گیرد.

[2] https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

[3] https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html#sklearn.tree.DecisionTreeClassifier.predict

[4] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html#sklearn.metrics.accuracy_score

[5] https://scikit-learn.org/stable/modules/tree.html#tree

[6] https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

۲ نکات پاسخ دهی

- تمرینات به صورت مرتب و خوانا بارگذاری شوند.
- برای تمرینات غیر عملی که به صورت تایپی ارسال شوند امتیاز تشویقی درنظر گرفته می‌شود.
- کد های خود را حتما در فایل PDF نیز قرار دهید.
- در سوالات توضیحی، قدرت تحلیل افراد ملاک مقایسهٔ پاسخ ها خواهد بود.
- فایل پایتون و یا Notebook برای تمرینات ضمیمه شود و همه به صورت یک فایل zip بارگذاری شوند. فایل zip را با فرمت DM4022_HW3_[StudentNumber].zip نام‌گذاری کنید.
- درصورت وجود ابهام خاص می توانید موارد را با دستیار آموزشی مطرح کنید.