آزمایشگاه تحلیل پیشرفته کلان داده (ابدال)

تمرین ۴: طبقه‌بندی–۱
(درخت تصمیم)

استاد: دکتر حسین رحمانی

اردیبهشت ۱۴۰۰

مقدمه

- مهلت ارسال تمرین تا ساعت ۲۳:۵۹ تاریخ ۱۴۰۰/۰۲/۳۱ است و قابل تمدید نخواهد بود.
- به ازای هر روز تاخیر ۲۵ درصد از نمره تمرین کسر خواهد شد.
- پاسخ به سوالات این تمرین باید در قالب یک گزارش با فرمت PDF ارائه شود.
- به همراه فایل گزارش تمرین، فایل کدهای اجراشده نیز پیوست شود.
- تمامی فایل‌های این تمرین (گزارش و کدها) در قالب یک فایل فشرده با نامگذاری زیر ارسال شود.
  StudentNumber_FirstName_LastName_HW4.zip
- فایل تمرین را از طریق سامانه LMS ارسال نمایید.
- رعایت نکات نگارشی در نوشتن گزارش نمره مثبت خواهد داشت.
- برای پاسخ به سوالات این تمرین حتما باید از زبان برنامه نویسی پایتون استفاده شود.

در این بخش سعی داریم به کمک زبان پایتون و با استفاده از scikit-learn به پیاده‌سازی درخت تصمیم بپردازیم و از آن برای پیش‌بینی اینکه آیا بیماری به بیماری قلبی مبتلا اســـت یا خیر کمک بگیریم. در ادامه مراحل انجام کار به ترتیب توضـیح داده شـده اسـت در هر مرحله خروجی، توضـیحات و تحلیل‌های لازم را به صورت کامل در گزارش ارائه دهید.

۱- پیش‌پردازش داده

در این تمرین از یکی از دیتاست‌های موجود در مخزن یادگیری ماشین UCI استفاده می‌کنیم. به طور خاص، ما قصد داریم از مجموعه بیماری های قلبی استفاده کنیم. این مجموعه داده به ما امکان می‌دهد که بر اساس جنسیت، سن، فشارخون و معیارهای دیگر فردی که به بیماری قلبی مبتلا شده است را پیش‌بینی کنیم.

لینک معرفی دیتاست:

https://archive.ics.uci.edu/ml/datasets/Heart+Disease

لینک دانلود دیتاست:

https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/processed.cleveland.data

۱-۱- وارد کردن داده

- داده‌ها را از فایل مورد نظر بخوانید و در dataframe ذخیره کنید
- نام ستون‌های را به صورت زیر وارد کنید

Columns= ['age' ,'sex' ,'cp' ,'restbp' ,'chol' ,'fbs' ,'restecg' ,'thalach' ,'exang' ,'oldpeak ,' 'slope' ,'ca' ,'thal' ,'hd']

- پیش‌پردازش‌های لازم را روی داده‌ها انجام دهید
- ستون مشخص‌کننده بیماری قلبی (ستون 'hd') را به عنوان ستونی انتخاب کنید که می‌خواهید پیش‌بینی کنید
- سایر ستون‌ها را به عنوان ویژگی‌هایی انتخاب کنید که برای انجام پیش‌بینی استفاده می‌شود

۱-۲- رمزگذاری One-Hot

- با استفاده از تابع "get_dummies()" ستون‌های غیربایتری را رمزگذاری One-Hot کنید

۳-۱- ساخت مدل درخت تصمیم

در ادامه به ساخت مدل درخت تصمیم می‌پردازیم. در هر مرحله اگر چندین حالت وجود دارد فرضیات خود را گزارش کنید و میزان بررسی و گزارش دقیق‌تر شما در نمره‌دهی شما لحاظ می‌شود.

۱-۳-۱- مقیاس‌بندی داده

- داده‌های تست و آموزش را جدا کنید (random_state=42)
- از تابع "scale()" استفاده کنید و داده‌ها آموزش و تست را مقیاس‌بندی کنید

۲-۳-۱- ساخت نمونه اولیه

- یک نمونه اولیه از درخت تصمیم با استفاده داده‌های آموزش و تست بسازید
- ساختار درختی آن را رسم کنید
- با استفاده از ماتریس درهم ریختگی (Confusion Matrix) و داده‌های تست عملکرد این درخت تصمیم را ارزیابی کنید

۳-۳-۱- هرس درخت تصمیم

- درخت تصمیم را با Cost complexity pruning هرس کنید
- به ازای آلفاهای متفاوت Accuracy را محاسبه کنید و نمودار آن را برای آموزش و تست در یک شکل رسم کنید بهترین نتیجه برای کدام مقدار آلفا است؟ (محور افقی آلفا و عمودی Accuracy باشد)
- به ازای الفاهای مختلف 5-fold cross validation اجرا کنید و میانگین و واریانسAccuracy را محاسبه و در نموداری این خروجی‌ها را نمایش دهید
- با رسم این نمودار بهترین مقدار آلفا را مشخص کنید

۴-۳-۱- ساخت، ارزیابی، رسم درخت تصمیم

- بر اساس بهترین مقدار آلفا درخت تصمیم جدیدی بسازید
- به ازای بهترین آلفا بدست آمده ماتریس درهم ریختگی (Confusion Matrix) را رسم کنید
- ساختار درختی بعد هرس شدن را رسم کنید