به نام خدا

سری دوم تمارین درس داده کاوی

استاد درس:

دکتر مریم امیر مزلقانی

نیم سال اول ۱۴۰۲-۱۴۰۳

راه ارتباطی:

Aut.DataMining.Fall@gmail.com

بخش عملی:

هدف از این بخش، پیاده سازی و مقایسه عملکرد مدل های مختلف دسته‌بندی می باشد. برای پیاده سازی این مدل ها، مجموعه داده ای درنظر گرفته شده است که هدف آن پیش بینی درآمد افراد با توجه به ویژگی های افراد در این مجموعه داده است. این مجموعه داده در پوشه Practical قرار داده شده است. مجموعه داده Adult_TrainDataset برای آموزش مدل ها در نظر گرفته شده است که حاوی۳۲۵۶۱ نمونه می باشد که در دو کلاس دسته بندی شده اند. در این مجموعه داده، ستون Income حاوی برچسب ها و مابقی ستون ها، ویژگی هایی هستند که بایستی برای پیاده سازی مدل، از آن ها استفاده نمایید.

- در این پروژه برای پیاده سازی مدل ها می توانید از مدل های آماده کتابخانه sklearn استفاده کنید.
- مجموعه داده های آموزش را به کمک حداقل دو مدل نمودار، بصری سازی کنید و به تحلیل آن ها بپردازید. با توجه به وجود مقادیر از دست رفته (Null) در این مجموعه داده، باید مقادیر از دست رفته را با مقادیر مناسب جایگزین کنید و یا ستون مربوطه را حذف نمایید.
- از آنجایی که برخی از ستون ها حاوی مقادیر Categorical هستند، در صورت استفاده از آنها برای آموزش مدل، باید مقادیر این ستون ها را با استفاده از روش های موجود، Encode کرده و به مقادیر عددی تبدیل کنید. توجه داشته باشید که روش های مختلفی جهت Encode کردن داده های Categorical وجود دارد که از جمله آنها می توان به One hot Encoding وLabel Encoding اشاره کرد. این دو روش را با هم مقایسه کرده و توضیح دهید با انتخاب کدام روش می توان به نتایج بهتری دست یافت و با روش انتخاب شده، به Encode کردن داده ها بپردازید.
- در صورت نیاز می توانید قبل از آموزش مدل ها، پیش پردازش های مختلفی همچون نرمال سازی، استانداردسازی و... را روی این مجموعه داده اعمال کنید.
- در آموزش هر مدل ، در صورت تنظیم هایپرپارامترهای مدل در کتابخانه sklearn، دلایل این انتخاب ها و تحلیل خودتان را از نتایج حاصل شده، گزارش دهید.
- در نهایت بعد از آموزش مدل ها، برای ارزیابی عملکرد مدل ها، از مجموعه داده Adult_TestDataset استفاده نمایید و با استفاده از ماتریس درهم ریختگی (confusion matrix) و معیارهای ارزیابی recall، precision ،accuracy وf1-score، به تحلیل نتایج بدست آمده بپردازید. توجه داشته باشید که برای سادگی کار، می توانید با استفاده از توابع آماده sklearn ، ماتریس درهم ریختگی و معیارهای ارزیابی را بدست آورید.
- توجه داشته باشید که ارایه گزارش هر بخش الزامی است و بایستی در فایل نوت بوک خود، بعد از هر سلول مربوط به کد، گزارش هر بخش را در یک سلول متنی بنویسید.