به‌نام خدا

فاز شناخت و فهم مسئله‌ی پروژه‌ی درس مبانی داده‌کاوی
1401-02

اعضای گروه:

سعید قاسمی 9826113
رسول کامکار 9826653

عنوان پروژه:

طبقه‌بندی درآمد افراد (بیشتر یا کمتر از $50,000) بر اساس اطلاعات آنها مانند سطح تحصیلات، نوع شغل، سن، نژاد و جنسیت

الف – مشکل چیست؟ (توضیح: چرا این مسئله یا داده تعریف شده است در واقع اگر این پروژه انجام نمی شد چه مشکلی، حل نشده باقی می ماند.)

بر اساس منبع اصلی دیتا https://archive.ics.uci.edu/ml/datasets/adult
این مسئله برای پیش‌بینی درآمد سرشماری (Census Income) افراد براساس مشخصات آن‌ها تعریف شده است
منظور از درآمد سرشماری، درآمد خالص سالیانه افراد بدون احتساب کسر مالیات است
از نمونه مسائلی که این پروژه می‌تواند حل کند، مقایسه آمار سرشماری با پیشبینی مدل و شناسایی موارد غیرعادی است که می‌تواند به کشف تقلب‌های مالی یا سرشماری‌های غلط منجرشود

ب-سوال داده کاوی و معیار ارزیابی آن چیست؟ (توضیح: قرار است چه مقدار(ستون یا اطلاعات) را پیش بینی (تخمین یا برآورد) کنید؟ و قصد دارید از چه اقلام اطلاعاتی برای این پیش بینی استفاده کنید؟معیار ارزیابی چیست و ایده آن چقدر است؟)

ستون هدف این پروژه income_>50K است
این ستون یک متغیر باینری است که نشان می‌دهد درآمد شخص بالاتر از 50000 است یا خیر
پیشبینی این ستون به صورت classification انجام می‌شود و از همه featureهای موجود به غیر از fnlwgt به دلیل correlation به شدت پایین برای اینکار استفاده می‌شود.
معیارهای ارزیابی پیشبینی ما F1 Score، Recall، Precision و جدول confusion matrix است

ج- مشخصات دیتاست و ویژگیهای موجود در آن

دیتاست شامل 14 ستون feature و یک ستون هدف است
ستون هدف income_>50K، یک متغیر binary است که بیشتر یا کمتر بودن درآمد فرد از 50000 را نشان می‌دهد

age: Integer (سن)
workclass: Categorical (دسته بندی شغل)
fnlwgt: Integer (توضیحات مربوط به این ستون در این لینک موجود است)

education: Categorical (سطح تحصیلات)
educational-num: Integer (کد شده ستون قبلی)
marital-status: Categorical (وضعیت ازدواج)
occupation: Categorical (شغل)
relationship: Categorical (وضعیت رابطه)
race: Categorical (نژاد)
sex: Categorical (جنسیت)
capital-gain: Integer (سود سرمایه مانند مسکن)
capital-loss: Integer (ضرر سرمایه به دلیل کاهش ارزش)
hours-per-week: Integer (ساعات کار در هفته)
native-country: Categorical (کشور محل تولد)

د- فعالیت پیش رو و بیان مختصر ایده تیم

در ابتدا با تکنیک‌های مختلف visualization، روابط بین ستون‌های مختلف داده مانند دسته‌بندی هر یک از کلاس‌ها بر اساس جنسیت، دسته بندی براساس درجه‌های تحصیلی مختلف، درآمد در بازه های سنی و... را مصورسازی کرده تا تصویر و فهمی کلی از دیتای موجود داشته باشیم
پس از آن مرحله پیش پردازش را بر روی دیتا انجام میدهیم، مانند حذف ستون‌های بیهوده، بازیابی یا حذف رکوردهای ناقص، کد کردن اطلاعات categorical و ...
در مرحله آخر، مدل‌های مختلف را بر روی اطلاعات موجود آموزش داده و عملکرد آن‌هارا بررسی و گزارش می‌کنیم