به نام خدا

دسته‌بندی درآمد افراد

پروژه درس داده‌کاوی

استاد: دکتر حمیدرضا حکیم‌داودی

رسول کامکار – سعید قاسمی

ارائه پروژه: اردیبهشت ۱۴۰۲

مقدمه

یکی از چالش‌های مهم در حوزه سرشماری و آمار جمعیت، تخمین درآمد افراد برای مقاصدی مانند دادن امتیازات خاص برای دسته‌های درآمدی خاص است. توانایی پیشبینی درآمد افراد می‌تواند کاربردهای بسیاری برای سیاست‌گذاران، جامعه شناسان و صاحبان کسب‌وکار باشد. به طور مثال، یک بیزینس می‌تواند بر اساس اطلاعات اجتماعی–اقتصادی مخاطبین خود درآمد آن‌ها را پیشبینی کرده و تبلیغات خود را بر افراد پردرآمد متمرکز کند.

در این دیتاست[1] ، اطلاعات مربوط به افراد مختلف ذخیره شده‌است. این اطلاعات شامل داده‌هایی مانند کشور، سطح تحصیلات، جنسیت و سن می‌شوند. این اطلاعات از Census database در سال ۱۹۹۶ گردآوری شده‌اند. هدف تشخیص و تفکیک افراد به دو دسته کم‌درآمد و پردرآمد (با مرز درآمد سالیانه ۵۰ هزار دلار) است.

در این گزارش، ما خلاصه‌ای از فعالیت‌های انجام شده در این پروژه از جمله بررسی و شناخت داده، پیش‌پردازش، انتخاب ویژگی‌ها و مدل‌های انتخاب شده و نتیجه آموزش آن‌ها ارائه شده است.

۱ روش‌ها

۱-۱ اطلاعات کلی دیتاست

دیتاست دارای حدود ۴۴ هزار رکورد و شامل ستون‌های زیر است:

جدول ۱: ویژگی‌های دیتاست

متغیر	نوع	توضیحات
age	کمی	سن شخص
workclass	کیفی	نوع شغل شخص
fnlwgt	کمی	وزن نهایی مالی که بر اساس شرایط شخص مقدار داده می‌شود[2]
education	کیفی	سطح تحصیلات شخص
educational-num	کمی	عدد متناظر با سطح تحصیلات (به ترتیب از سطوح پایین به بالا)
marital-status	کیفی	وضعیت تأهل شخص
occupation	کیفی	شغل شخص
relationship	کیفی	وضعیت رابطه شخص
race	کیفی	نژاد
gender	کیفی	جنسیت
capital-gain	کمی	میزان سود
capital-loss	کمی	میزان ضرر
hours-per-week	کمی	ساعات کار در هفته
native-country	کیفی	کشوری که شخص بومی آنجاست
income_>50K	کمی	متغیر هدف. اینکه آیا درآمد شخص بیشتر از ۵۰هزار دلار است یا خیر

همچنین تعدادی مقادیر `null` وجود دارد:

workclass	occupation	native-country
2498	2506	763

از آنجا که حدود ۲۵۰۰ رکورد هم برای `workclass` و هم `occupation` مقدار `NaN` دارند، و چون همه رکوردهای حاوی مقدار `NaN` درصد کمی از کل دیتاست را تشکیل می‌دهند، تصمیم به حذف آنها گرفته شد.

۱-۲ تبدیل ستون‌های کیفی

از آنجا که `education-num` مقادیر عددی متناظر با `education` را در بر دارد، ستون `education` برای ساخت مدل حذف می‌شود. `gender` نیز چون تنها دو مقدار دارد تبدیل به مقادیر ۰ و ۱ می‌شود. بقیه متغیرهای کیفی نیز به کمک روش MCA[3] کدگذاری می‌شوند.

۱-۳ انتخاب ویژگی‌های مناسب

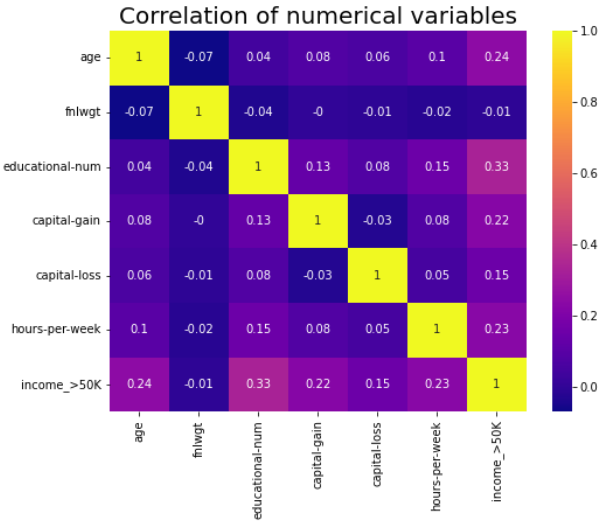
برای حذف ویژگی‌های بی‌تاثیر، دیتاست دو معیار `f-score` و `mutual information` با روش `SelectKBest` ارزیابی می‌شوند تا هر کدام ۳ ویژگی بی‌تاثیر را مشخص کنند تا اشتراکات آنها از دیتاست حذف شود.

۱-۴ نرمال‌کردن داده‌ها

برای نرمال‌کردن داده‌ها از روش `MinMax` استفاده شده‌است تا توزیع داده‌ها حفظ شود. همچنین از آنجا که مدل‌های انتخابی `Decision Tree` و `KNN` هستند، نیازی به `Binning` نیست. از طرفی به دلیل مقاومت این مدل‌ها درمقابل داده پرت و نویز (با پیش هرس برای `DTree` و به ازای `K` به اندازه کافی بزرگ برای `KNN`)، از حذف این داده‌ها صرف‌نظر شده‌است.

۲ آزمایشات

در ابتدا، برای بدست‌آوردن شهودی از ارتباط بین ویژگی‌ها، ماتریس همبستگی بررسی می‌شود:



شکل ۱: میزان همبستگی ویژگی‌های کمی

همانطور که در شکل ۱ مشاهده می‌شود، میزان همبستگی داده‌ها به طور کلی بالا نیست و بیشترین همبستگی بین متغیر هدف و میزان تحصیلات است که مقدار ۰.۳۳ دارد. این اعداد با وجود اینکه در مقیاس ۰ تا ۱ خیلی بزرگ نیستد، به دلیل تعداد بسیار بالای رکوردها تا حد خوبی قابل اعتماد و معتبر هستند. از طرفی `fnlwgt` همبستگی نزدیک به صفر با همه‌ی دیگر ویژگی‌ها دارد.

۲-۱ پیش‌پردازش

برای تبدیل داده‌های کیفی به مقادیر کمی به روش MCA، از کتابخانه Prince[4] استفاده شده تا این ستون‌ها در ۳ ستون کمی خلاصه شوند.

برای حذف ستون‌های بی‌تاثیر، پس از اعمال روش‌های ذکرشده، با معیار f-score سه ویژگی MCA_1، MCA_2 و fnlwgt و با معیار mutual information، سه ویژگی capital-loss، fnlwgt و gender به عنوان ویژگی‌های بی‌تاثیر تعیین شدند. از آنجا که fnlwgt در هردو وجود دارد، این ویژگی از دیتاست حذف می‌شود.

در نهایت داده‌ها به نسبت ۸۵،۱۵ برای آموزش و تست تقسیم می‌شوند.

۲-۲ مدل‌سازی

برای شروع، **baseline** مشخص می‌شود؛ با توجه به توزیع متغیر هدف، ساده‌ترین مدل که جدای از ورودی، مقدار صفر می‌دهد، دقت ۷۵٪ دارد که دقت پایه در نظر گرفته می‌شود.

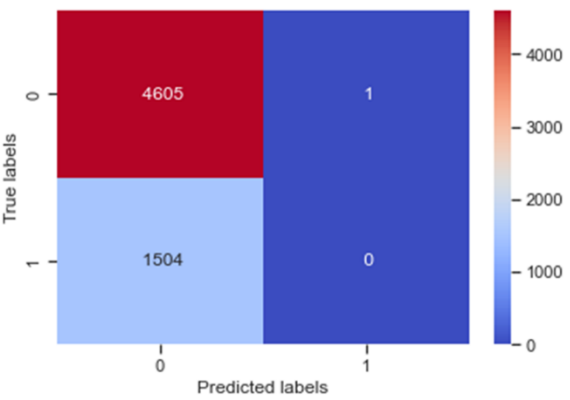
در ادامه مدل **RandomForest** با CrossValidation (5-fold) آموزش داده‌می‌شود. از آنجا که فرآیند آموزش و تنظیم هایپرپارامترها زمان‌بر بود، مدل پس از آموزش و تنظیم در یک فایل ذخیره می‌شود تا برای استفاده‌های بعدی آماده باشد. در نهایت بهترین هایپرپارامترهای بدست‌آمده عبارتند از:

criterion	min_samples_left	min_samples_split	max_depth	max_features	n_estimators
Gini	1	2	4	log2	264

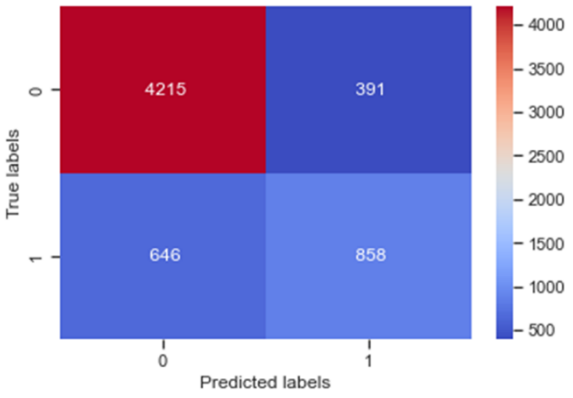
مدل **KNN** نیز برای انتخاب مقدار بهینه K ارزیابی می‌شود. طبق ارزیابی، مقدار بهینه K برابر ۱۵ است.

۳ نتایج

در نهایت از داده‌ی تست برای ارزیابی نهایی مدل‌ها استفاده می‌شود. confusion matrix مدل‌ها به شکل زیر است:



شکل ۳: Random Forest Confusion Matrix



شکل ۲: KNN Confusion Matrix

جدول ۲: نتایج ارزیابی مدل‌ها

	Accuracy	F-score(0)	F-score(1)
RandomForest	0.748	0.856	0
KNN	0.83	0.89	0.542

۴ جمع‌بندی

طبق نتایج، اینگونه برداشت می‌شود که RandomForest در چنین شرایطی به خوبی عمل نمی‌کند. از آنجا که داده‌های کیفی در سه ویژگی کمی خلاصه شده‌اند و درخت تصمیم نیز در هر راس تنها یک ویژگی را ارزیابی می‌کند، عدم در نظر گرفتن همزمان ویژگی‌های MCA_x می‌تواند دلیل عملکرد ضعیف این مدل باشد. بنابراین انتخاب مدل جایگزین و یا کدگذاری‌های دیگر میتوانند از جمله راهکارهایی برای تحلیل و بررسی بیشتر روی این دیتاست باشند.

در پایان از مدل نهایی می‌توان برای کشف تقلب و فرار مالیاتی و یا خطای سرشماری در داده‌های آماری بهره برد.

۵ مراجع

[1] "Income Dataset," [Online]. Available: https://archive.ics.uci.edu/ml/datasets/adult.

[2] "fnlwgt," [Online]. Available: https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.names.

[3] "prince," [Online]. Available: https://maxhalford.github.io/prince/mca/.

[4] "MCA," [Online]. Available: https://www.ibm.com/docs/he/spss-statistics/25.0.0?topic=categories-multiple-correspondence-analysis.