

تکلیف فصل ششم

در یک موضوع داده‌کاوی در ارتباط با کلاهبرداری اینترنتی روی داده‌های تراکنشی نتیجه بدین شرح است. ۸۸ رکورد کلاهبرداری وجود داشته که ۳۰ رکورد آن به‌درستی رده‌بندی شده است. ۹۵۰ رکورد عدم کلاهبرداری وجود داشته که ۹۲۰ رکورد آن به‌درستی رده‌بندی شده است.

الف: ماتریس ارزشیابی مربوط به این موضوع را تشکیل دهید.

ب: معیارهای دقت (Accuracy)، نرخ خطا (Error Rate)، حساسیت (Sensitivity) یا (Recall)، خاص بودن (Specificity)، درستی (Precision) و F1-score را برای این داده‌ها محاسبه نمایید.

پ: با توجه به نتایج به‌دست‌آمده چه تفسیری در مورد این نتایج دارید؟

پاسخ قسمت الف:

اگر ارزش کلاهبرداری را P در نظر بگیریم، ۸۸ رکورد کلاهبرداری وجود داشته است که از این تعداد ۳۰ رکورد به‌درستی کلاهبرداری تشخیص داده شده و ۵۸ رکورد به‌غلط عدم کلاهبرداری تشخیص داده شده است؛ بنابراین:

$$TP = 30, FN = 58$$

همچنین ۹۵۰ رکورد عدم کلاهبرداری وجود داشته است که از این تعداد ۹۲۰ رکورد به‌درستی عدم کلاهبرداری تشخیص داده شده و ۳۰ رکورد به‌غلط کلاهبرداری تشخیص داده شده است؛ بنابراین:

$$FP = 30, TN = 920$$

درنتیجه ماتریس ارزشیابی به شکل زیر ترسیم می‌گردد:

	کلاهبرداری P	عدم کلاهبرداری N
کلاهبرداری P	TP = 30	FN = 58
عدم کلاهبرداری N	FP = 30	TN = 920

پاسخ قسمت ب:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} = \frac{30 + 920}{30 + 30 + 58 + 920} = \frac{950}{1038} = 0.915$$

$$\text{Error Rate} = \frac{FP + FN}{TP + FP + FN + TN} = \frac{30 + 58}{30 + 30 + 58 + 920} = \frac{88}{1038} = 0.085$$

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{30}{30 + 58} = \frac{30}{88} = 0.340$$

$$\text{Specificity} = \frac{TN}{TN + FP} = \frac{920}{920 + 30} = \frac{920}{950} = 0.968$$

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{30}{30 + 30} = \frac{30}{60} = 0.500$$

$$\text{F1 Score} = \frac{2 * \text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}} = \frac{2 * 0.340 * 0.500}{0.340 + 0.500} = \frac{0.34}{0.84} = 0.405$$

- مقدار ۹۰٪-ای Accuracy به این معنی است که از هر ۱۰ برچسب ۱-ای نادرست است و ۹ مورد صحیح است.
- با توجه به Error Rate ۸/۵٪ از پیش‌بینی‌های ما اشتباه است.
- مقدار ۳۴٪-ای Recall به این معنی است که تقریباً از هر ۱۰ فرد کلاه‌بردار، تقریباً ۷ نفر در برنامه ما فراموش می‌شوند و ۳ نفر به‌عنوان کلاه‌بردار برچسب زده می‌شوند.
- مقدار ۹۷٪-ای Specificity به این معنی است که از هر ۱۰ فرد غیر کلاه‌بردار، تقریباً ۱ نفر به‌عنوان کلاه‌بردار (در بدترین حالت) و ۹ نفر به‌درستی به‌عنوان غیر کلاه‌بردار برچسب‌گذاری شده‌اند (اگرچه احتمال اینکه تمامی ۱۰ نفر غیر کلاه‌بردار تشخیص داده شوند بسیار بیشتر است).
- مقدار ۵۰٪-ای Precision به این معنی است که به‌طور متوسط از هر ۱۰ فرد کلاه‌بردار که توسط برنامه ما برچسب‌گذاری شده است، ۵ نفر غیر کلاه‌بردار و ۵ نفر کلاه‌بردار هستند.
- F1-score میانگین هارمونیک درصد امتیاز یک آزمون در هر یک از معیارهای Precision و Recall است.