پاسخ تمرین سری ۱

محمدرضا عزیزی
۹۸۱۳۱۰۲۲
دانشکده مهندسی کامپیوتر
دانشگاه صنعتی امیرکبیر (پلی‌تکنیک تهران)
mrazizi@aut.ac.ir

۱ راهکاری برای پیش‌پردازش داده‌ها

داده‌ها با استفاده از کتابخانه pandas بارگذاری کرده و از آن‌جایی که ستون‌های داده‌ها نام ندارد، ابتدا با توجه به توضیحات دیتاست، یک لیست برای نام ستون‌ها تعریف کرده و این نام‌ها را به ستون‌های دیتافریم بارگذاری شده اضافه می‌کنیم. لیست این نام‌ها به ترتیب برابر است با:

```
["age", "workclass", "fnlwgt", "education", "education-num",
"marital-status", "occupation", "relationship", "race", "sex",
"capital-gain", "capital-loss", "hours-per-week", "native-country", "label"]
```

ستون آخر که مربوط به برچسب داده‌است با با نام label نام‌گذاری کرده‌ایم.

سپس یک دیکشنری تعریف می‌کنیم که هر دوتایی کلید/مقدار آن، خود یک دیکشنری است. برای مثال به ازای کلید label، دیکشنری زیر را داریم:

```
'label': {'<=50K':0, '>50K': 1}
```

به ازای تمامی مقادیر تمامی ستون‌هایی که مقدار عددی ندارند، این دیکشنری را تعریف کرده و به هر مقدار اسمی، یک عدد نسبت می‌دهیم. این اعداد از برای هر متغیر از ۰ شروع شده و یک واحد یک واحد افزایش می‌یابد.

در نهایت با استفاده از تابع replace از کتابخانه pandas، طبق دیکشنری تعریف شده، مقادیر اسمی را به مقادیر عددی تبدیل می‌کنیم.

۲ بارگذاری داده‌ها و انجام پیش‌پردازش

	age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	label
0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
3	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
4	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K

	age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	label
0	39	5	77516	0	13	2	8	3	0	1	2174	0	40	0	0
1	50	1	83311	0	13	0	4	2	0	1	0	0	13	0	0
2	38	0	215646	3	9	1	6	3	0	1	0	0	40	0	0
3	53	0	234721	2	7	0	6	2	4	1	0	0	40	0	0
4	28	0	338409	0	13	0	5	0	4	0	0	0	40	12	0