РК 1

РТ5-61Б Слкуни Герман

Задание 3, датасет 3

Для заданного набора данных произведите масштабирование данных (для одного признака) и преобразование категориальных признаков в количественные двумя способами (label encoding, one hot encoding) для одного признака.

Импортирование нужных библиотек

In [33]:
```
import pandas as pd
from sklearn.preprocessing import MinMaxScaler, LabelEncoder, OneHotEncoder
```

Загрузка данных

In [34]:
```
data = pd.read_csv('toy_dataset.csv')
data.head()
```

Out[34]:

	Number	City	Gender	Age	Income	Illness
0	1	Dallas	Male	41	40367.0	No
1	2	Dallas	Male	54	45084.0	No
2	3	Dallas	Male	42	52483.0	No
3	4	Dallas	Male	40	40941.0	No
4	5	Dallas	Male	46	50289.0	No

Масштабирование признака Age

In [35]:
```
column = 'Age'
scaler = MinMaxScaler()
scaled_column = scaler.fit_transform(data[[column]])
data[f'{column}_scaled'] = scaled_column
data.head()
```

Out[35]:

	Number	City	Gender	Age	Income	Illness	Age_scaled
0	1	Dallas	Male	41	40367.0	No	0.400
1	2	Dallas	Male	54	45084.0	No	0.725
2	3	Dallas	Male	42	52483.0	No	0.425
3	4	Dallas	Male	40	40941.0	No	0.375
4	5	Dallas	Male	46	50289.0	No	0.525

Преобразование категориального признака City в количественый

In [36]:
```
category_column = 'City'
print(f'Уникальных значений категориального признака {category_column}: {ler
```

```
Уникальных значений категориального признака City: 8
```

Label Encoding

In [37]:
```
label_encoder = LabelEncoder()
encoded_data = label_encoder.fit_transform(data[category_column])
data[f'{category_column}_label_encoded'] = encoded_data
data.head()
```

Out[37]:

	Number	City	Gender	Age	Income	Illness	Age_scaled	City_label_encoc
0	1	Dallas	Male	41	40367.0	No	0.400	
1	2	Dallas	Male	54	45084.0	No	0.725	
2	3	Dallas	Male	42	52483.0	No	0.425	
3	4	Dallas	Male	40	40941.0	No	0.375	
4	5	Dallas	Male	46	50289.0	No	0.525	

One Hot Encoding

In [38]:
```
onehot_encoder = OneHotEncoder(sparse_output=False, drop='first')
encoded_data = onehot_encoder.fit_transform(data[[category_column]])
onehot_columns = [f'{category_column}_{cat}' for cat in onehot_encoder.categ
onehot_df = pd.DataFrame(encoded_data, columns=onehot_columns, index=data.ir
data = pd.concat([data, onehot_df], axis=1)
data.head()
```

Out[38]:

	Number	City	Gender	Age	Income	Illness	Age_scaled	City_label_encoc
0	1	Dallas	Male	41	40367.0	No	0.400	
1	2	Dallas	Male	54	45084.0	No	0.725	
2	3	Dallas	Male	42	52483.0	No	0.425	
3	4	Dallas	Male	40	40941.0	No	0.375	
4	5	Dallas	Male	46	50289.0	No	0.525	