UNIVERSIDADE FEDERAL DA FRONTEIRA SUL

APRENDIZAGEM SUPERVISIONADA:
CUSTOMER SHOPPING TRENDS

Componente Curricular: Inteligência Artificial
Professor(a): Felipe Grando
Acadêmico(a): Matheus Vieira Santos

BASE DE DADOS

- ➔ Customer Shopping Trends Dataset: Journey into Consumer Insights and Retail Evolution with Synthetic Data do Kaggle;
- ➔ Esse conjunto de dados oferece informações interessantes sobre o comportamento dos consumidores e seus padrões de compra;
- ➔ É importante mencionar que este conjunto de dados é composto por dados sintéticos. Ele foi criado pelo ChatGPT para ser utilizado em estudos de análise de dados e machine learning. Além disso, é relevante considerar que sua estrutura foi feita simulando uma experiência real de compra, ou seja, espelhando o mundo real. Isso, por sua vez, faz com que o modelo apresentado, embora agora seja feito com dados sintéticos, mais tarde, se desejado, pode ser aplicado a dados reais, revelando informações concretas.

BASE DE DADOS

➔ O conjunto possui 3900 registros e as seguintes features: Identificador do usuário, Idade, Gênero, Item Comprado, Categoria, Valor da Compra, Localização, Tamanho, Cor, Estação, Avaliação, Possui Assinatura, Tipo de Frete, Desconto Aplicado, Cupom Aplicado, Realizou Compras Anteriormente, Método de Pagamento e Frequência de Compras.

Customer ID	Age	Gender	Item Purchased	Category	Purchase Amount (USD)	Location	Size	Color	Season	Review Rating	Subscription Status	Shipping Type	Discount Applied	Promo Code Used	Previous Purchases	Payment Method	Frequency of Purchases
2896	56	Female	Hoodie	Clothing	86	Montana	L	Green	Summer	4.60	No	Standard	No	No	29	Bank Transfer	Monthly
2752	27	Female	Dress	Clothing	52	Minnesota	S	Indigo	Fall	3.10	No	Free Shipping	No	No	50	Venmo	Monthly
1224	69	Male	Pants	Clothing	24	Kansas	L	Red	Winter	3.90	No	Free Shipping	Yes	Yes	21	Bank Transfer	Weekly
2485	60	Male	Hoodie	Clothing	97	New Hampshire	M	Green	Summer	4.80	No	2-Day Shipping	No	No	50	Cash	Every 3 Months
3286	58	Female	Hat	Accessories	31	Hawaii	XL	Magenta	Fall	4.60	No	Free Shipping	No	No	11	Cash	Weekly

OBJETIVO COM MODELO

- ➔ Modelo classificador: Árvores de Decisão;
- ➔ Classificação Multiclasse;
- ➔ Desconsiderando os itens da categoria Acessórios;
- ➔ Com as seguintes features: 'Age', 'Gender', 'Purchase Amount (USD)', 'Location', 'Size', 'Color', 'Season', 'Frequency of Purchases', 'Review Rating' e 'Payment Method';
- ➔ A ideia é classificar a categoria do item de acordo: 'Clothing' 'Footwear' e 'Outerwear'.

LIMPEZA DOS DADOS

- ➔ Desconsiderar as features: 'Customer ID', 'Item Purchased', 'Subscription Status' e 'Promo Code Used';
- ➔ Dado que o conjunto de dados é sintético, não foi necessário realizar mais tratamentos de limpeza de dados.

```
How many NaN values by attribute:
Age                       0
Gender                    0
Category                  0
Purchase Amount (USD)     0
Location                  0
Size                      0
Color                     0
Season                    0
Review Rating             0
Shipping Type             0
Discount Applied          0
Previous Purchases        0
Payment Method            0
Frequency of Purchases    0
dtype: int64
How many 0 values by attribute:
Age                       0
Gender                    0
Category                  0
Purchase Amount (USD)     0
Location                  0
Size                      0
Color                     0
Season                    0
Review Rating             0
Shipping Type             0
Discount Applied          0
Previous Purchases        0
Payment Method            0
Frequency of Purchases    0
dtype: int64
```

TRANSFORMAÇÃO DOS DADOS

➔ Reduzindo cores em cores claras, escuras e coloridas

```
def categorize_color(color):
    light_colors = {'Beige', 'Cyan', 'Gold', 'Lavender', 'Peach', 'Pink', 'Silver', 'White', 'Yellow'}
    dark_colors = {'Black', 'Brown', 'Charcoal', 'Gray', 'Maroon', 'Olive', 'Purple'}
    colorful_colors = {'Blue', 'Green', 'Indigo', 'Magenta', 'Orange', 'Red', 'Teal', 'Turquoise',
'Violet'}

    if color in light_colors:
        return 'Light'
    elif color in dark_colors:
        return 'Dark'
    elif color in colorful_colors:
        return 'Colorful'

df['Color'] = df['Color'].apply(categorize_color)
```

TRANSFORMAÇÃO DOS DADOS

➔ Transformando propriedades com o LabelEncoder

Age => [18 19 … 69 70]

Gender => ['Female' 'Male']

Category => ['Clothing' 'Footwear' 'Outerwear']

Location => ['Alabama' 'Alaska' … 'Wisconsin' 'Wyoming']

Color => ['Colorful' 'Dark' 'Light']

Season => ['Fall' 'Spring' 'Summer' 'Winter']

Size => ['L' 'M' 'S' 'XL']

Frequency of Purchases => ['Annually' …]

Purchase Amount (USD) => [20 21 … 99 100]

Shipping Type => ['2-Day Shipping' 'Express' …]

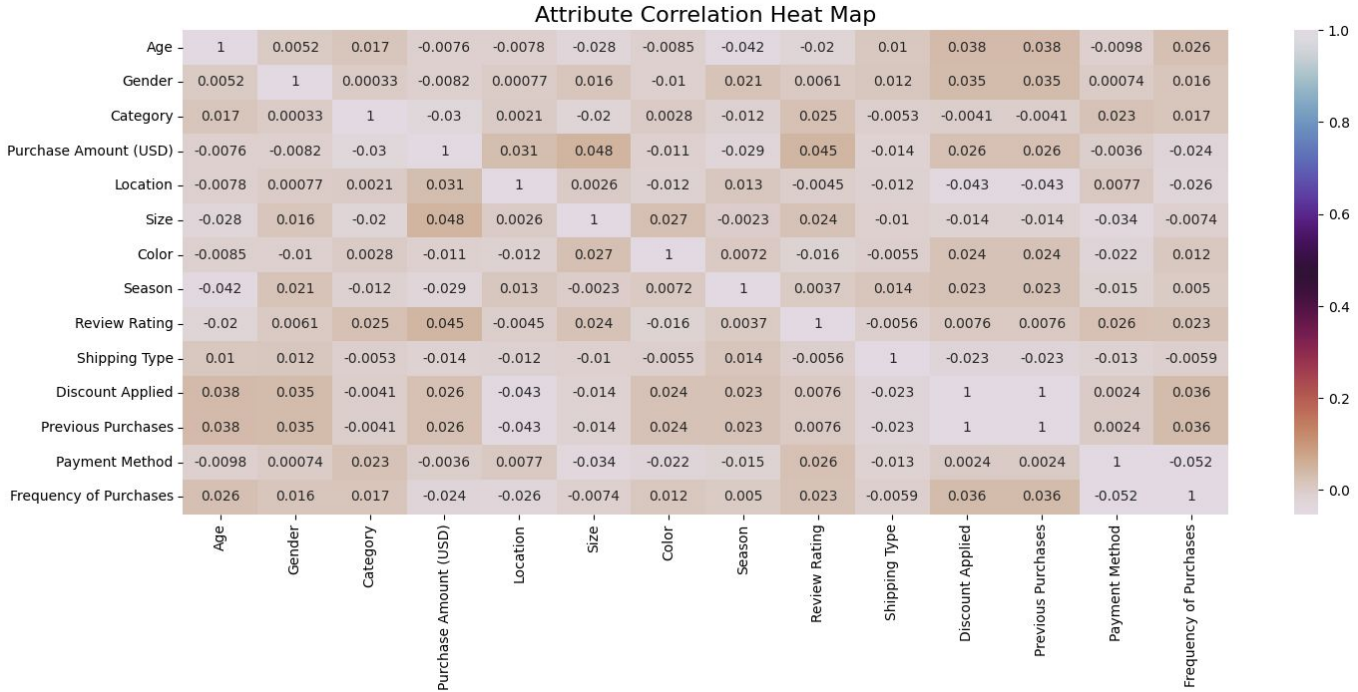
Review Rating => [2.5 2.6 … 4.9 5.]

Discount Applied => ['No' 'Yes']

Previous Purchases => [1 2 … 49 50]

Payment Method => ['Bank Transfer' 'Cash' 'Credit Card' 'Debit Card' 'PayPal' 'Venmo']

COMPREENDENDO OS DADOS



NORMALIZAÇÃO DOS DADOS

➔ Nesse passo preparamos os dados para algoritmos de machine learning, assim garantimos que todas as variáveis contribuam igualmente para o modelo, evitando que variáveis com escalas maiores dominem aquelas com escalas menores.

```
colnames = ['Age','Gender','Purchase Amount (USD)','Location','Size',
            'Color','Season','Frequency of Purchases','Review Rating','Payment
Method'];pp.MinMaxScaler()
X[colnames] = scaler.fit_transform(X[colnames])
```

APLICANDO O ALGORITMO

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=145)

print("0: ",(y_train==0).sum())
print("1: ",(y_train==1).sum())
over_sampler = RandomOverSampler(random_state=100)
X_train, y_train = over_sampler.fit_resample(X_train, y_train)

print("After:")
print("0: ",(y_train==0).sum())
print("1: ",(y_train==1).sum())

clf = RandomForestClassifier(max_depth=20, min_samples_leaf=1, min_samples_split=2,
n_estimators=250)
clf.fit(X_train, y_train)
```

MÉTRICAS

- **F1 score:** 0.5244733318005639
- **Accuracy:** 0.6416040100250626
- **Precision:** 0.603099538840053
- **Recall:** 0.6416040100250626

Actual \ Predicted			
	494	9	4
	188	7	1
	94	0	1

REFERÊNCIAS

BANERJEE, S. "Customer Shopping Trends Dataset". Disponível em: https://www.kaggle.com/datasets/iamsouravbanerjee/customer-shopping-trendsdataset. Acesso em 04 de junho de 2024.