Classification problem

Classify whether an email is spam or not spam based on its content

\sunsun-datateathyme

Import modules

```
import pandas as pd
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import train_test_split
```

Sample data

```
## sample data (0 = not spam, 1 = spam)
emails = pd.DataFrame({
    'word_count': [150, 250, 80, 500, 120],
    'contains_link': [0, 1, 0, 1, 0],
    'is_spam': [0, 1, 0, 1, 0]
})
emails
```

	word_count	contains_link	is_spam
0	150	0	0
1	250	1	1
2	80	0	0
3	500	1	1
4	120	0	0

Next steps: Generate code with `emails` | View recommended plots | New interactive sheet

Prepare data

```
## prepare data
X = emails[['word_count', 'contains_link']] # Features
y = emails['is_spam'] # Target variable
```

Split data

```
## split data
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

Train model

```
## train model
model = DecisionTreeClassifier()
model.fit(X_train, y_train)
```

```
DecisionTreeClassifier()
```

Prediction

```
# Predict a new email
new_email = [[300, 1]] # High word count, contains a link
prediction = model.predict(new_email)
```

```
/usr/local/lib/python3.11/dist-packages/sklearn/utils/validation.py:2739: UserWarning: X does not have valid feature names, but Deci
  warnings.warn(
```

Print result

```
## print result
print(f"Prediction: {'Spam' if prediction[0] == 1 else 'Not Spam'}")
```

```
Prediction: Spam
```

```
Start coding or generate with AI.
```