4_Features

January 14, 2025

[ ]:
```
'''
   In machine learning and data analysis, features are the individual␣
↪measurable properties
   or characteristics of the data that are used as inputs to a model. Each␣
↪feature represents
   an attribute of the data, and collectively, features are used to make␣
↪predictions or perform analysis.

   For example, in a dataset of house prices, features could include -->

   House Size (sq ft) : Numerical feature.
   Number of Bedrooms : Numerical feature.
   Location : Categorical feature.
'''
```

[ ]:
```
'''
   In pandas, iloc is a function that allows you to select data by index␣
↪positions (row/column numbers).
   It is particularly useful for selecting specific rows and columns in a␣
↪DataFrame, which is useful when
   selecting features or specific parts of a dataset.
'''
```

[ ]:
```
#   You can use iloc to select specific features (columns) based on their index␣
↪positions.
#   df.iloc[row_indices, column_indices]
#   row_indices: Specifies the row(s) you want to select.
#   column_indices: Specifies the column(s) (features) you want to select.
```

[1]:
```
import numpy as np
import pandas as pd
```

[2]:
```
data = pd.read_csv('Data/Data.csv')
data
```

[2]:

	Country	Age	Salary	Purchased
0	France	44.0	72000.0	No

```
1    Spain  27.0  48000.0       Yes
2  Germany  30.0  54000.0        No
3    Spain  38.0  61000.0        No
4  Germany  40.0      NaN       Yes
5   France  35.0  58000.0       Yes
6    Spain   NaN  52000.0        No
7   France  48.0  79000.0       Yes
8  Germany  50.0  83000.0        No
9   France  37.0  67000.0       Yes
```

[5]:
```
#   Selecting a Single Feature (Column)

#   Select the 'Country' column (index 0)
#   Select all rows but only the first column

country = data.iloc[:, 0]
print(country)
```

```
0     France
1      Spain
2    Germany
3      Spain
4    Germany
5     France
6      Spain
7     France
8    Germany
9     France
Name: Country, dtype: object
```

[6]:
```
#   Selecting Multiple Features (Columns)

#   Select 'House Size' and 'Bedrooms' columns (indices 0 and 1)
#   Select columns 0 and 1

selected_features = data.iloc[:, [0, 1]]
print(selected_features)
```

```
   Country   Age
0   France  44.0
1    Spain  27.0
2  Germany  30.0
3    Spain  38.0
4  Germany  40.0
5   France  35.0
6    Spain   NaN
7   France  48.0
8  Germany  50.0
```

```
9   France  37.0
```

[7]:
```
#   Selecting All Features Except the Target

#   Select all features except 'Price' (index 2)
#   Select all

x_data = data.iloc[:, :-1].values
```

[9]:
```
x_data
```

[9]:
```
array([['France', 44.0, 72000.0],
       ['Spain', 27.0, 48000.0],
       ['Germany', 30.0, 54000.0],
       ['Spain', 38.0, 61000.0],
       ['Germany', 40.0, nan],
       ['France', 35.0, 58000.0],
       ['Spain', nan, 52000.0],
       ['France', 48.0, 79000.0],
       ['Germany', 50.0, 83000.0],
       ['France', 37.0, 67000.0]], dtype=object)
```

[8]:
```
#   Selecting Target

#   Select target

y_data = data.iloc[:, -1].values
```

[10]:
```
y_data
```

[10]:
```
array(['No', 'Yes', 'No', 'No', 'Yes', 'Yes', 'No', 'Yes', 'No', 'Yes'],
      dtype=object)
```

[ ]:
```
#   We use .values() or .to_numpy() to get an np-array
#   As Imputation works on arrays
```