Assignment PWS 21 March

Q1. What is the difference between Ordinal Encoding and Label Encoding? Provide an example of when you might choose one over the other.

Ordinal Encoding:

- Encodes categorical variables with an inherent order into numeric values.
- Maintains the order in the data.
- Example: Education levels (High School = 1, Bachelor's = 2, Master's = 3).

Label Encoding:

- Assigns unique numeric values to each category without assuming any order.
- Used for nominal data with no inherent hierarchy.
- Example: Fruits (Apple = 0, Banana = 1, Cherry = 2).

Example Choice:

- Use **Ordinal Encoding** for ordered categories like "Low, Medium, High".
- Use **Label Encoding** for unordered categories like "Red, Blue, Green".

Q2. Explain how Target Guided Ordinal Encoding works and provide an example of when you might use it in a machine learning project.

Target Guided Ordinal Encoding:

- Assigns numeric values to categories based on their relationship with the target variable.
- Categories are ordered by the mean (or median) of the target variable within each category.

Example:

- A dataset with a "City" column and a target variable "Sales":
    - Compute the average sales for each city.
    - Assign ranks based on these averages (e.g., City A = 3, City B = 2, City C = 1).

Use Case:

- Use this technique when categorical variables are strongly correlated with the target variable, such as customer segments influencing purchase amounts.

Q3. Define covariance and explain why it is important in statistical analysis. How is covariance calculated?

Covariance:

- Measures the degree to which two variables change together.
- Indicates the direction of the relationship between variables (positive or negative).

Importance:

- Helps understand the relationship between variables.
- Identifies features that may be predictive of the target variable in machine learning.

Calculation: For two variables X and Y:

$$\text{Cov}(X, Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

Q4. For a dataset with the following categorical variables: Color (red, green, blue), Size (small, medium, large), and Material (wood, metal, plastic), perform label encoding using Python's scikit-learn library. Show your code and explain the output.

Code:

```
from sklearn.preprocessing import LabelEncoder

import pandas as pd

# Dataset

data = pd.DataFrame({

    'Color': ['red', 'green', 'blue'],

    'Size': ['small', 'medium', 'large'],

    'Material': ['wood', 'metal', 'plastic']
```

```
})

# Applying Label Encoding

encoder = LabelEncoder()

data_encoded = data.apply(encoder.fit_transform)

print(data_encoded)
```

Output:

Color	Size	Material
2	2	2
1	1	0
0	0	1

Explanation:

- Each category is encoded as an integer (e.g., red = 2, green = 1, blue = 0).
- Label Encoding is applied column-wise.

Q5. Calculate the covariance matrix for the following variables in a dataset: Age, Income, and Education Level. Interpret the results.

Assume the dataset:

Age	Income	Education Level
25	40000	2
30	50000	3
35	60000	4

Calculation (using Python):

```
import numpy as np

import pandas as pd

# Data

data = pd.DataFrame({

    'Age': [25, 30, 35],
```

```
    'Income': [40000, 50000, 60000],
    'Education Level': [2, 3, 4]
})

# Covariance Matrix
cov_matrix = data.cov()
print(cov_matrix)
```

Output:

	Age	**Income**	**Education Level**
Age	25.0	50000.0	2.5
Income	50000.0	100000000	50000.0
Education Level	2.5	50000.0	0.5

Interpretation:

- Positive covariances (e.g., Age-Income) indicate a direct relationship.
- Larger values suggest stronger relationships.

Q6. You are working on a machine learning project with a dataset containing several categorical variables, including "Gender" (Male/Female), "Education Level" (High School/Bachelor's/Master's/PhD), and "Employment Status" (Unemployed/Part-Time/Full-Time). Which encoding method would you use for each variable, and why?

1. **Gender**:
   - **Binary Encoding**: Male = 0, Female = 1.
   - Justification: Binary categorical data.
2. **Education Level**:
   - **Ordinal Encoding**: High School = 1, Bachelor’s = 2, Master’s = 3, PhD = 4.
   - Justification: Inherent order in the levels.

3. **Employment Status**:
    - **One-Hot Encoding**: Creates binary columns for each category.
    - Justification: No inherent order.

Q7. You are analyzing a dataset with two continuous variables, "Temperature" and "Humidity," and two categorical variables, "Weather Condition" (Sunny/Cloudy/Rainy) and "Wind Direction" (North/South/East/West). Calculate the covariance between each pair of variables and interpret the results.

Steps:

1. **Encode Categorical Variables**:
    - Apply Label Encoding to "Weather Condition" and "Wind Direction."
2. **Calculate Covariance**:
    - Use the covariance formula for continuous variables.
    - Use software tools like Python or R for calculation.

Code:

```
from sklearn.preprocessing import LabelEncoder

import pandas as pd

# Data

data = pd.DataFrame({

   'Temperature': [30, 35, 40],

   'Humidity': [70, 65, 60],

   'Weather Condition': ['Sunny', 'Cloudy', 'Rainy'],

   'Wind Direction': ['North', 'South', 'East']

})

# Encoding Categorical Variables

encoder = LabelEncoder()

data['Weather Condition'] = encoder.fit_transform(data['Weather Condition'])

data['Wind Direction'] = encoder.fit_transform(data['Wind Direction'])
```

```
# Covariance
cov_matrix = data.cov()
print(cov_matrix)
```

Interpretation:

- High covariance values indicate strong relationships.
- Zero or near-zero values suggest weak or no relationships.