**A picture containing shape, arrow

Description automatically generatedDummy Variables**

Instructions:

Please share your answers filled inline in the word document. Submit code files wherever applicable.

Please ensure you update all the details:

**Name: Vishvash C**

**Batch Id: 23012024**

**Topic: Data Pre-Processing**

**Problem Statement:**

Data is one of the most important assets. It is often common that data is stored in distinct systems with different formats and forms. Non-numeric form of data makes it tricky while developing mathematical equations for prediction models. We have the preprocessing techniques to make the data convert to numeric form. Explore the various techniques to have reliable uniform standard data, you can go through this link:

<https://360digitmg.com/mindmap-data-science>

1. Prepare the dataset by performing the preprocessing techniques, to have all the features in numeric format.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Index | Animals | Gender | Homly | Types |
| 1 | Cat | Male | Yes | A |
| 2 | Dog | Male | Yes | B |
| 3 | Mouse | Male | Yes | C |
| 4 | Mouse | Male | Yes | C |
| 5 | Dog | Female | Yes | A |
| 6 | Cat | Female | Yes | B |
| 7 | Lion | Female | Yes | D |
| 8 | Goat | Female | Yes | E |
| 9 | Cat | Female | Yes | A |
| 10 | Dog | Male | Yes | B |

**Hints:**

For each assignment, the solution should be submitted in the below format.

1. Work on each feature to create a data dictionary as displayed in the image displayed below:
2. Refer to the animal\_category.csv data set.
3. Research and perform all possible steps for obtaining the solution.
4. All the codes (executable programs) should execute without errors.
5. Code modularization should be followed.
6. Each line of code should have comments explaining the logic and why you are using that function.

import pandas as pd

df = pd.read\_csv(r"C:/Users/Lenovo/Downloads/Study material/EDA/InClass\_DataPreprocessing\_datasets/Animal\_category.csv")

### Identify duplicate records in the data ###

# Duplicates in rows

# help(data.duplicated)

# duplicate = data.duplicated() # Returns Boolean Series denoting duplicate rows.

# duplicate

# sum(duplicate)

# Create dummy variables

df\_new = pd.get\_dummies(df)

df\_new.columns

df\_new.info()

df\_new\_1 = pd.get\_dummies(df, drop\_first = True)

Index Animals\_Dog Animals\_Goat Animals\_Lion Animals\_Mouse Gender\_Male Homly\_Yes Types\_B Types\_C Types\_D Types\_E

0 1 False False False False True True False False False False

1 2 True False False False True True True False False False

2 3 False False False True True True False True False False

3 4 False False False True True True False True False False

4 5 True False False False False True False False False False

from sklearn.preprocessing import OneHotEncoder

# Creating instance of One-Hot Encoder

enc = OneHotEncoder() # initializing method

enc\_df = pd.DataFrame(enc.fit\_transform(df.iloc[:, 1:]).toarray())

# Get the feature names from the OneHotEncoder

feature\_names = enc.get\_feature\_names\_out(input\_features=df.columns[1:])

# Assign feature names to the encoded DataFrame

enc\_df.columns = feature\_names

enc\_df.columns

enc\_df.info()

Animals\_Cat Animals\_Dog Animals\_Goat Animals\_Lion Animals\_Mouse Gender\_Female Gender\_Male Homly\_No Homly\_Yes Types\_A Types\_B Types\_C Types\_D Types\_E

0 1.0 0.0 0.0 0.0 0.0 0.0 1.0 0.0 1.0 1.0 0.0 0.0 0.0 0.0

1 0.0 1.0 0.0 0.0 0.0 0.0 1.0 0.0 1.0 0.0 1.0 0.0 0.0 0.0

2 0.0 0.0 0.0 0.0 1.0 0.0 1.0 0.0 1.0 0.0 0.0 1.0 0.0 0.0

3 0.0 0.0 0.0 0.0 1.0 0.0 1.0 0.0 1.0 0.0 0.0 1.0 0.0 0.0

4 0.0 1.0 0.0 0.0 0.0 1.0 0.0 0.0 1.0 1.0 0.0 0.0 0.0 0.0

import pandas as pd

from sklearn.preprocessing import LabelEncoder

# Define the label encoder

encoder = LabelEncoder()

# Fit label encoder and transform the "Types" column

df['Encoded\_Types'] = encoder.fit\_transform(df['Types']) + 1 # Adding 1 to start encoding from 1

df.head()

Index Animals Gender Homly Types Encoded\_Types

0 1 Cat Male Yes A 1

1 2 Dog Male Yes B 2

2 3 Mouse Male Yes C 3

3 4 Mouse Male Yes C 3

4 5 Dog Female Yes A 1

data = pd.DataFrame({

'Types': ['A', 'B', 'C', 'C', 'A', 'B', 'D', 'E']

})

# Define the mapping of values to alphabets

value\_mapping = {'A': 4, 'B': 2, 'C': 8, 'D': 6, 'E': 10} # Add more alphabets if needed

# Map values to the "Types" column

data['Encoded\_Types2'] = data['Types'].map(value\_mapping)

print(data)

Types Encoded\_Types2

0 A 4

1 B 2

2 C 8

3 C 8

4 A 4

5 B 2

6 D 6

7 E 10

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of Feature** | **Description** | **Type** | **Relevance** |
| **ID** |  | **Quantitative/ Nominal** | **Irrelevant (ID does not provide useful information)** |
| Index | Unique identifier for each record | Nominal | Relevant |
| Animals | Type of animals present | Nominal | Relevant |
| Gender | Gender of the individuals | Nominal | Relevant |
| Homly | Indicator for whether the animal is homely or not | Nominal | Relevant |
| Types | Different types/categories | Nominal | Relevant |