**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: Shabezade Srikanth**

**Batch Id: DS 07102022**

**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 the 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 |

import pandas as pd

import numpy as np

data1=pd.read\_csv('F:/Assignments/Assignment 4 data preprocessing/DataSets/animal\_category.csv')

data1.dtypes

data1.info

data1.describe

data2=pd.get\_dummies(data1)

data1.drop(['Index'],axis=1,inplace=True)

from sklearn.preprocessing import OneHotEncoder

enc= OneHotEncoder(handle\_unknown='ignore')

enc\_data1=pd.DataFrame(enc.fit\_transform(data1).toarray())

enc\_data1

from sklearn.preprocessing import LabelEncoder

labelencoder= LabelEncoder()

x=data1.iloc[:,0:4] #input

y=data1['Types'] #output

data1.columns

x['Animals']=labelencoder.fit\_transform(x['Animals'])

x['Types']=labelencoder.fit\_transform(x['Types'])

y=labelencoder.fit\_transform(y)

y=pd.DataFrame(y)

data2=pd.concat([x,y],axis=1)

data2.columns

data2=data2.rename(columns={0:'Type1'})

data2

**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 animal\_category.csv data set.
3. Research and perform all possible steps for obtaining 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

**A picture containing shape, arrow

Description automatically generatedGrading Guidelines:**

**Note: 1. An Assignment submission is considered complete only when successful executable code(s), and documentation explaining the applied solution and results are provided. Failing to submit either of them will be considered an invalid submission and will not be considered for evaluation.**

**2. Assignments submitted after the deadline date will affect your grades.**

**Grading:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Ans** | **Date** |  |  | **Ans** | **Date** |
| Correct | On time | A | 100 |  |  |
| 80% & above | On time | B | 85 | Correct | Late |
| 50% & above | On time | C | 75 | 80% & above | Late |
| 50% & below | On time | D | 65 | 50% & above | Late |
|  |  | E | 55 | 50% & below |  |
| Copied/No Submission |  | F | 45 |  |  |

* **Grade A: (>= 90):** When all assignments are submitted on or before the given deadline date
* **Grade B: (>= 80 and < 90):** 
  + When assignments are submitted on time but less than 80% of questions asked in assignments are completed. (or)
  + All assignments were submitted, however, after the given deadline
* **A picture containing shape, arrow

  Description automatically generatedGrade C: (>= 70 and < 80):** 
  + When assignments are submitted on time but less than 50% of questions asked in assignments are completed. (or)
  + Less than 80% of questions asked in assignments are submitted after the deadline
* **Grade D: (>= 60 and < 70):** Assignments submitted after the Deadline and with 50% or less of questions
* **Grade E: (>= 50 and < 60):** 
  + Less than 30% of questions asked in the assignments are submitted after the deadline (OR)
  + Less than 30% of questions asked in the assignments are submitted before deadline

**Grade F: (< 50):** Copied submission or No submission