**Zero-Variance Features**

Instruction

Please ensure you update all the details:

**Name: Vishvash C**

**Batch Id: 23012024**

**Topic: Data Pre-Processing**

Variance measures how far a set of data is spread out. A variance of zero indicates that all the data values are identical. There are various techniques to remove this for transforming the data into the suitable one for prediction.

**Problem statement:**

Find which columns of the given dataset with zero variance, and explore various techniques used to remove the zero variance from the dataset to perform certain analysis.



**Hints:**

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

1. Work on each feature of the dataset to create a data dictionary as displayed in the below image:



1. Consider the Z\_dataset.csv dataset.
2. Research and perform all possible steps for obtaining the solution.
3. All the codes (executable programs) should execute without errors.
4. Code modularization should be followed.
5. 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/Z\_dataset.csv")

df.dtypes

# If the variance is low or close to zero, then a feature is approximately constant and will not improve the performance of the model.

# In that case, it should be removed.

ds = df.iloc[:,1:5]

ds

ds.columns

for column in ds.columns:

range = max(ds[column]) - min(ds[column])

print(range)

for a in ds.columns:

print(f"the variance of {a} is ", df[a].var())

#since variance of square.breadth is near zero we will drop the column

df.drop(columns=["square.breadth"], inplace=True)

df['square.length'].var()

#colour has equal number of categories so we will retain it

print(df)

df.colour.value\_counts()

**Output:**

3.6000000000000005

2.4000000000000004

5.9

2.4

the variance of square.length is 0.6856935123042505

the variance of square.breadth is 0.1899794183445188

the variance of rec.Length is 3.1162778523489942

the variance of rec.breadth is 0.5810062639821029

Id square.length rec.Length rec.breadth colour

0 1 5.1 1.4 0.2 Blue

1 2 4.9 1.4 0.2 Blue

2 3 4.7 1.3 0.2 Blue

3 4 4.6 1.5 0.2 Blue

4 5 5.0 1.4 0.2 Blue

.. ... ... ... ... ...

145 146 6.7 5.2 2.3 Orange

146 147 6.3 5.0 1.9 Orange

147 148 6.5 5.2 2.0 Orange

148 149 6.2 5.4 2.3 Orange

149 150 5.9 5.1 1.8 Orange

[150 rows x 5 columns]

Out[85]:

colour

Blue 50

Green 50

Orange 50

Name: count, dtype: int64

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of Feature** | **Description** | **Type** | **Relevance** |
| **ID** |  | **Quantitative/ Nominal** | **Irrelevant (ID does not provide useful information)** |
| Id | Unique identifier for each square | Nominal | Irrelevant |
| square.length | Length of the square | Quantitative | Relevant |
| square.breadth | Breadth of the square | Quantitative | Relevant |
| rec.Length | Length of the rectangle | Quantitative | Relevant |
| rec.breadth | Breadth of the rectangle | Quantitative | Relevant |
| colour | Colour of the shape | Nominal | Irrelevant |