**Zero - Variance Features**

Instruction

Please ensure you update all the details:

**Name: ANANDAKRISHNAN K V**

**Batch Id: 19042021**

**Topic: Preliminaries for Data Analysis**

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, explore various techniques used to remove the zero variance from the dataset to perform certain analysis.



**Ans:-**

**Python code:-**

############ Zero variance #############

import pandas as pd

### import data set "Z\_datasetcsv" as dataframe ##

df = pd.DataFrame(Z\_datasetcsv)

df1 = df.copy(deep=True) ### keeping original dataset by creating a copy df1

df1.columns ## column name

#drop Id Column it doesn't gonna give any data information

df1.drop(['Id'], axis = 1, inplace=True)

df1.dtypes

### here the last column is a string data or character data. So for conducting zero variance checking applying label encoding to get corresponding numeric data

from sklearn.preprocessing import LabelEncoder

## creating instance of label encoder

labelencoder = LabelEncoder()

df1["colour"]= labelencoder.fit\_transform(df1["colour"])

###### zero variance operation ###

df1.shape

## importing ###

from sklearn.feature\_selection import VarianceThreshold

# Feature selector that removes all low-variance features that meets the variance threshold limit

var\_thres = VarianceThreshold(threshold=0.2) # Threshold is subjective.

var\_thres.fit(df1) ### fit the var\_thres to data set df1

# Generally we remove the columns with zero variance, but i took thresold value 0.2 (Near Zero Variance)

var\_thres.get\_support() ### it giving an array out, where zero variant column treat as False value. we already fit var\_thres to df1. so it gives corresponding information on df1

df1.columns[var\_thres.get\_support()] ## non-zero variant column names

constant\_columns = [column for column in df1.columns if column not in df1.columns[var\_thres.get\_support()]]

print(len(constant\_columns)) ### number of zero variant variables

for feature in constant\_columns:

print(feature) ### names of corresponding zero variant columns

df2 = df1.drop(constant\_columns, axis = 1) ### data set with non-zero variant variables or features

df2.colour = df.colour ### since colour is not-zero variant column, replace back to its original format

df2["Id"] = df.Id ### adding column back to data set

### coclusion ##

print("final data set with non\_zero variance feature: ",df2)

Hints**:**

A picture containing shape, arrow

Description automatically generatedFor 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 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