CHANAKYA
UNIVERSITY

Analyzing Artistic Patterns: A Comparative Study of Classification and Clustering Techniques on the Artist Dataset.

Author: Sohan Singh Rawat

Reg no: CU22MSD004A

School of Mathematics and Natural Science

Submitted to Prof. Usha Subramanian

Introduction:

The project is about the Artist which contains the data about the artist and the work done by the artist. The dataset contains columns that can be used for the analysis. It contains 15343 rows of data and there are a total of 8 columns that give the idea about the artist. It has columns ConstituentID, Display Name, ArtistBio, Nationality, Gender, Begin Date, EndDate, and Count of Work.

Project Work:

In this, there is 3 section data Cleaning, Data Preprocessing, and Data modelling.

- Firstly, I have checked the Missing values using Python and I got to know that the dataset contains missing values in 3 columns. Since to tackle the missing we can either use mode for categorical since I feel to drop the rows with missing values. After dropping the rows, the new dataset has about 11420 data.
- Secondly, Here the column Artist Bio has values such as Nationality with Born date year, and Death year, and it is written together so I have separated them and added a new column for it. I have created Nationality and Born and End year Columns out of it.
- In the Gender columns, there was a silly mistake of Uppercase and lowercase and I have rectified it by using Python. Since I cannot use categorical values for the ML algorithm, I converted them into numerical values.
- Duplicate values are being removed.

All the preprocessing work is done and the Dataset is ready for use the Algorithms Here two algorithm is used Classification and Clustering for Classification I have used a Decision tree and in Clustering I have K means.

Using some sample and done in Excel:

Here I have taken some sample datasets and used a decision tree. So, I added extra column labels and based on the count of work I have distributed them. So, the artist below 100 low, more than 100 and less than 500 medium, and more than 500 High. Again, going deep, I got the insight there is only Two Female in the High category and both are dead, and out of 27 males two of them are still alive. In the low category, most of the artists are there and in these 3 third gender people are there and all of them are alive. In the medium category, there is no other gender and 84 artists are there and most of them are alive and done good about of work.

Similarly, I used this with Kmeans and took some random values from the dataset subsequently taking the distance from the data point and based on the distance cluster is assigned with the column having minimum value.

After analyzing I used 3 clusters and out of it only 25 clusters belonged to c2 and c3 clusters rest belonged to c1 The same process was done using Python in Jupyter Notebook using two techniques decision tree and Kmeans.

Outliers' detection is done in this dataset using histogram For this I have used 11421 datasets and bins ranging from 1 to 5050 and used Excel to get the output as shown in figure 1.

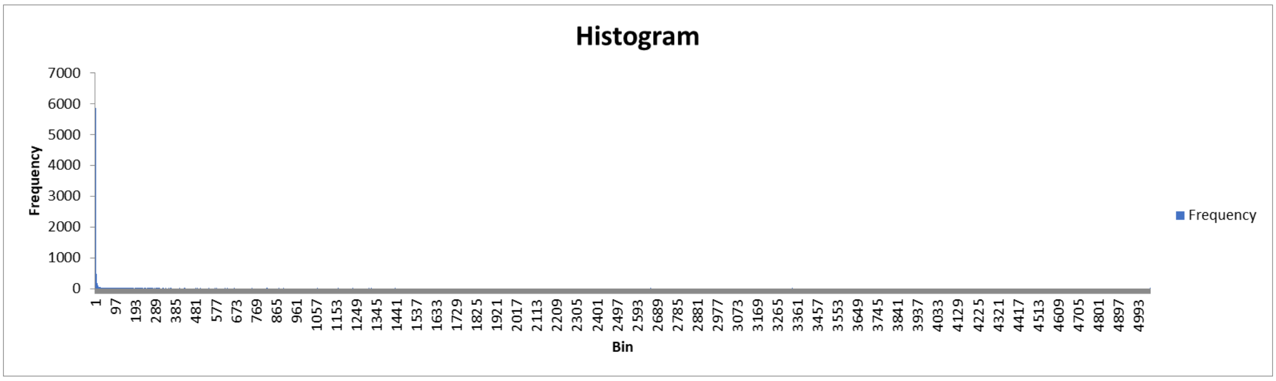

Figure 1: Outlier Detection

Here After visualizing the histogram it can be seen that there are some bars in the 1 which is big compared to the rest which gives the essence of the outliers., it is visible that outliers are there in the dataset.

Conclusion:

After analysis, I got the insight that Kmeans is better for this dataset because it is less complex than a decision tree since the decision tree is sensitive and the data is overfitting. A decision tree is good for seeing how Data is using various features for instance here first I have a Count of work and labelled them, then after that based on gender, and after that on the live status I did the analysis. Kmeans will be the better technique in terms of accuracy. In K means when k= 3 then it gives a better cluster with higher accuracy using Silhouette Score.