GOA COLLEGE OF ENGINEERING

"Bhausaheb Bandodkar Technical Education Complex"

Experiment No: 2 **Date:**

Lab Session 2: Data Pre-processing using Orange

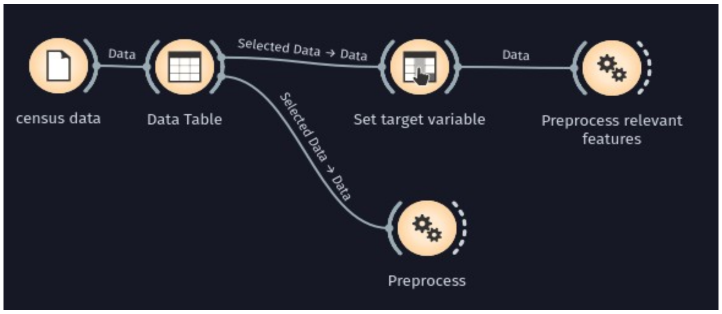
Aim: Perform data cleaning and pre processing on the given dataset

Problem Description: Perform the following on the given dataset

- Impute Missing values
- Discretize continuous values
- Continuize discrete values
- Select relevant attributes
- Select random attributes

Widgets Used:
Preprocess

Data Workflow:



Conclusion: Justify the strategies that you have used to achieve the aim of this experiment.

1) **Imputation:**
If we impute using most frequent/average value statistically most of the missing values will be filled in with the correct value. This is not true in case the absence of the value means something.

2) **Discretisation:**
I've Discretised using equal width binning and 10 bins. Thus we can still somewhat infer relative size between bins

3) **Continuization:**
I've Continuized categorical variables by setting the most frequent value as base. This would maintain the frequency information in the category.