Case Intern DSC Batch 1 – Theory

Name: Marvin Sachio

NIM : 2602064475

1. Explain briefly about Exploratory Data Analysis! Why we must do it before advanced to build a predictive model ?
   Answer : Exploratory Data Analysis in short EDA, is a process to analyze and summarize the dataset, this process will make us understand the dataset, which dataset that got pattern, outlier, or unclean, unclear, unfill data (missing data). By this process we could say that we detecting anomalies from the dataset that given. EDA also include with visualization, for better visual of the data distribution.
   Why we must do EDA before build model? The dataset that given might be unclear and unclean so we need to clean it to make more effective and cleaner yet accurate data for predictive model
2. What does it mean by imbalanced dataset? Give an example about imbalanced dataset problem!
   Answer : Imbalance dataset, a dataset given that the distribution of the value is not balance or similar in count, for example the data distribution given 80% and 20% of the binary value, this distribution is unbalance and could make the model biased to the data that is majority which is 80%.
   Example :
   Given the data from a survey about cancer, many participants of the survey that input 'No' to aware of cancer than 'Yes' this make unbalanced data that could, make biased model.

3. What method you should do when you handling a missing value on a numeric dataset and categorical dataset?
   Answer : handling missing value on a numeric dataset, I usually use central tendencies of data it self, inputing missing value with median, mode, or mean. Its central tendencies got own pro and cons but we can fill missing values with it. Handling missing value on a categorical dataset, can be filled with mode which is the data that appear the most, but if the data is missing too much mode gonna make the data too artificial.

   For some case, if the data missing about 70% from total data i chose to removing the category, this action is based on the data set will be over artificial if we fill it with central tendencies.

Verified by

Rafael Nicholas Tanaja as Data Science Club Leader on March' 12 2023.