MACHINE LEARNING PROJECT PHASE 1

DATA PREPROCESSING

Submission Date: ~~20th October 2020~~ 13 th Sep 2021

Note : Each student in the group should complete the phase 1 for their respective datasets pertaining to the project chosen. Marks will be awarded individually after viva/presentation.

Data Pre-processing :

- Numerical data
  - Normalization
  - Standardization
  - Imputing Missing values
  - Discretization
- Text data
  - Text related preprocessing (stop words removal, word stemming)
  - Conversion of text data to numeric (Tf-idf),
    Note: [Use python nltk package – eg: vectorization]
- Image Data
  - If color based processing, RGB components of each pixel can be feature vectors. You can also convert the image to other color spaces like YCbCr, HSV etc. Use CvtColor function in opencv python
  - Or color image can be converted to gray scale image using CvtColor
  - If needed, gray scale image can be converted to binary image using cv2.threshold function.
  - For any more complex recognition problems specific feature extraction methods like shape,texture,color etc may be extracted to get the feature vector.

- Data Summarization:
  - Use statistical methods to understand the data and apply the required methods
- Data Visualization:
  - Visualize the data using various plots like scatterplot, histograms, box plot etc and record your interpretations with varying values
- Data Interpretation:
  - Record all your findings and summary about data
  - Document all the results with relevant screenshots.
  - Upload the original as well as the cleaned data.