Data Intake Report

Name: APS Failure from Scania trucks
Report date: 2021-10-25
Internship Batch: NLP01
Version: 1.0
Data intake by: Juan Carlos Gutiérrez
Data intake reviewer: NA
Data storage location: https://archive.ics.uci.edu/ml/machine-learning-databases/00421/

Tabular data details:

File name: aps_failure_training_set.csv

Total number of observations	60000
Total number of files	1
Total number of features	171
Base format of the file	.csv
Size of the data	44.7 MB

File name: aps_failure_test_set.csv

Total number of observations	16000
Total number of files	1
Total number of features	171
Base format of the file	.csv
Size of the data	11.9 MB

Proposed Approach:

The goal is to have a definitive dataset, ready to analyse and/or use for modelling purposes. These are the steps I followed:

1. Review the data files: values, variables, dimension, etc.
2. Look for null values: in this case, those are represented by the string “na” in the data frames.
3. Encode the target from “neg” and “pos”, to 0 and 1.
4. Have the correct data types: integer for the target variable and float for the rest.
5. Standardize the values, since there are some outliers, I preferred to standardize instead of normalize, expecting to have a better scaling before modelling.
6. Work on the standardized data sets, given the variability of the different variables.