1. **DATASET**

We have used the NASA Metrics Data Program (MDP) datasets for prediction of software faults. Faults can occur in software even when the development is finished due to various human errors and shortcomings of the system used for software development. These faults can cause failure in the actual implementation of the software at the user’s end. The existing software fault prediction systems have done a helpful job in making error-free software. The systems are still unable to achieve high accuracy and robust quality check. Due to the ever-expanding scope of software, a system which can predict the faultiness of software with maximum accuracy and minimum manual interaction is mandatorily required.

The MDP dataset is a standard dataset prepared by NASA through collection of software data from various sources. It is a binary dataset which describes software based on various numerical measures. In other words, it contains the attributes or features which describe software, and a binary class which labels any instance of software as faulty or not faulty. This dataset is widely used all over the world for making software fault prediction systems and conducting research in this field. Before applying prediction models on the dataset, it is important to do a quality check. As the original dataset, which was made available by the organization, struggled with its quality due to a number of issues, such as duplicate data points, missing values, etc, it was necessary for researchers to build preprocessing algorithms for cleaning of data. Due to its heavy use and demand of the researchers, the MDP dataset evolved through time and it is now available in a preprocessed and cleaned state, which we have used in our experiments. This cleaned data is completely suitable to use with any prediction technique without any preprocessing. The cleaned MDP dataset is described in Table 1.

The validity of the preprocessing algorithm used to clean the dataset is highly important. It helps us in deciding if the data is suitable for using with various prediction models or not. We have used the cleaned data based on the quality issues described in the paper “Data Quality: Some Comments on the NASA Software Defect Datasets.”

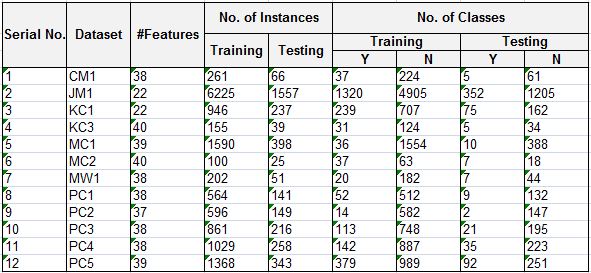


Table : *The table shows the description of the cleaned and preprocessed dataset from the NASA Metrics Data Program. The dataset have been divided into two parts- for training and testing.*

It removes the duplicate data points from the dataset and adds the missing data values.