Exploratory data analysis and processing

While examining the distribution of the features, we remark that there is only one categorical feature *PRI\_jet\_num = {0,1,2,3}* which is closely related with the fact that there are a lot of Nan values in the data. Indeed, checking the description of the features, we learn that a jet is a type of pseudo particle that is created as a result of the collision of other particles. We also realise that there are a few features which simply cannot be defined if there is only one jet or none.

Therefore, we decide to categorise the data in four different subsets so that some features can be removed depending on whether or not they are defined for the corresponding number of jets. Moreover, in each subset there are also Nan values in the feature *DER\_mass\_MMC* and we replace them by the corresponding mean. In conclusion, we end up with four data-sets without any Nan values.

|  |  |  |
| --- | --- | --- |
| Features | Jet 0 | Jet 1 |
| *DER\_deltaeta\_jet\_jet* | *Undef.* | *Undef.* |
| *DER\_mass\_jet\_jet* | *Undef.* | *Undef.* |
| *DER\_lep\_eta\_centrality* | *Undef.* | *Undef.* |
| *PRI\_jet\_leading\_pt* | *Undef.* |  |
| *PRI\_jet\_leading\_eta* | *Undef.* |  |
| *PRI\_jet\_leading\_phi* | *Undef.* |  |
| *PRI\_jet\_subleading\_pt* | *Undef.* | *Undef.* |
| *PRI\_jet\_subleading\_eta* | *Undef.* | *Undef.* |
| *PRI\_jet\_subleading\_phi* | *Undef.* | *Undef.* |

Undefined features in the data set per jet and per features