

Predictive analysis using big data technologies.

Jesús M. Antoñanzas, Alex Carrillo.

Advanced Databases, GCED, UPC - December 17th 2019

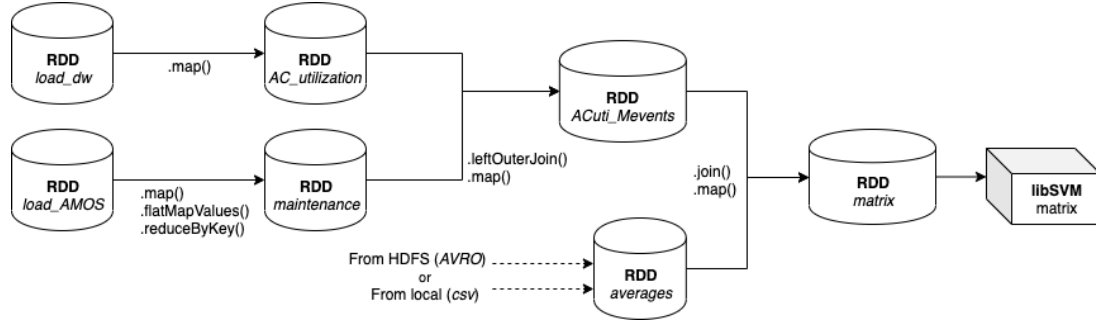


Figure 1: Data management pipeline schema.

Data management. Broadly, the pipeline makes use of *AMOS* and *aircraftutilization* to get the required metrics and create the response variable. Some nuances must be noticed; The creation of the 6 dates before each unscheduled maintenance has been implemented using generators, which lowers memory usage. We also consider the current date as positive with respect to the response variable (an unsch. maint. event happened sometime in the next 7 days). The *date* attribute has been replaced from `datetime.date()` format to `String` to ease type consistency. The *matrix* rdd has been converted to `LabeledPoint` type, as the *LibSVM* saving requires a (label, [features]) vector.

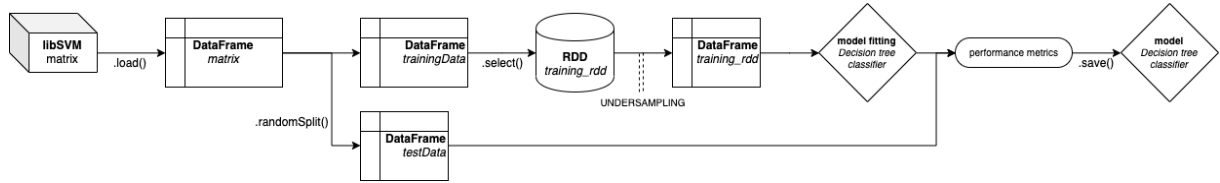


Figure 2: Data analysis pipeline schema.

Data analysis. When fitting the Decision Tree classifier, the training set is undersampled, as it is the most reasonable approach to deal with the low presence of unscheduled maintenances. After several trial and test implementations (a validation split was used), say, the 60/40% proportion of scheduled/unscheduled maintenances, respectively, seemed work well in general cases. As a way to validate the model *accuracy* and *weighted recall* have been considered.

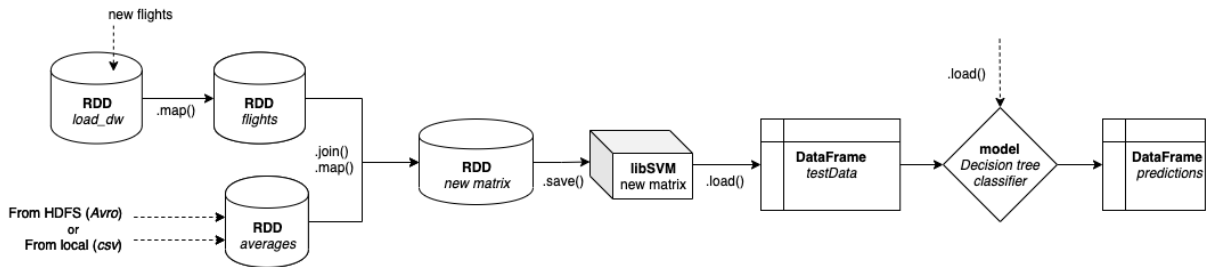


Figure 3: Data classifier pipeline schema.

Data classifier. Lastly, the data classifier pipeline reproduces the same flow from the beginning of the first pipe to the end of the second, but with some particularities. As you only want to take into account the new flights introduced to the DB, *AMOS* is not used. The new test data matrix must be created, converted and placed into the previously saved model. Those predictions are saved in a text file in case they may be looked up. We consider that one would want to predict unsch. maintenances for, say, the last week, so we implemented a way to input the date from which to start predicting.