Lachlan Sinclair

1/26/20

CS_419_400

Week 3 assignment 1: Handling Missing Feature Values in a Dataset

Reflections:

A. Comparing the way we have been using the two modelers and the results they produce it is clear that Watson Studio is much more suited for handling huge amounts of data. It also has a far nicer UI for generating multiple tables, SPSS just spam opens everything which is inconvenient. SPSS seems like it was developed to handle small data sets that are relatively easy to manually navigate. I would say the original vision of SPSS based of my experience so far was to create a light weight easy to use ML platform that can quickly generate proof of concepts using small datasets. This is clearly different than the big data oriented and deployment solution supported Watson Studio. Watson Studio was envisioned as being a useful tool implemented as a cloud service that utilizes big data and seamlessly integrates with the rest of their suite of tools. The process of selecting the which data asset to use and the clean formatting of the output tables highlights these differences.

B. I think overall the way we implemented filtering and partitioning on this small dataset could be used in far larger datasets without issue. Since this is a relatively small dataset, we could reasonably manually curate the data, if we were to do something like that it clearly would not scale to large data sets. But using a derive node and a select node to deal with missing values would work for any size of data, also the partitioning node doesn't seem to concern itself with the data size. This means there would likely be no additional work required to use this filtering and partitioning method on larger datasets, however I do suspect there would be some additional steps required to feed huge data sets into this stream.