Review for: Wytze Bekker’s “EDA Breast Cancer Proteomes: Dividing breast cancer patients into separate sub-classes.”

Reviewed by: Dennis Scheper (373689)

Your introduction looks great; you describe what EDA is about and which sources you are using. It also is compactly written and is clear in its description. However, you could add some abbreviations for files you are using instead of putting the whole filename in. I noticed this several times and found that it makes your EDA harder to read; introducing abbreviations would certainly improve this. Besides, I would advise you only to cite a reference in your text and state the whole reference in a source list, even if it is only one source you need to mention. It creates a nice and more familiar structure comparable to other EDAs and articles. You state that your main goal is to remove useless data. On top of that, you could briefly explain how you intend to accomplish this by, for example, mentioning to look at the variation between genes. It is a minor detail but creates a more understanding view for readers to state this in an introduction. The proposed research question seems very appropriate for a machine learning project.

Looking at your codebook, you very clearly show what each attribute is and visualizes, but not every description is filled in. If this is for a reason, you could mention this inside the table with an asterisk (*) or explain it somewhere in your article, or, if possible, fill in the missing ones. Descriptions that are filled in are self-explanatory—good job on that. You could also mention an attribute’s data type in the codebook. It seems not that necessary but is a nice addition to the information you are presenting.

The way you are describing the process of omitting NAs is very understanding. I can follow your explanations well enough to reproduce the same results. Looking at the section where you talk about the effects of omitting every NA record, you could explain which components are in a dimension (the first element are the records, and the second element represents the attribute name). You could also achieve this by defining labels above the dimension via the function cat().

Moving on to the usability section, I found how you incorporated multiple boxplots of the expressions of genes genius. This way, you visualize the outliers and range of valuation very well. You state that outliers are useful for machine learning. However, this is not true, as machine learning algorithms are sensitive to the range and distribution of attribute values. This way, data outliers can mislead the training process, which leads to less accurate models and, finally, poorer results. I think it is good to look into normalization techniques that remove outliers so they cannot disturb your conclusion. Normalization techniques that come to mind are linear scaling, log scaling, clipping, and Z-scores. I would not recommend a log scaling as your data is already presented in a log format. However, the rest seem completely open to me as your data is mostly scoring of expression data. I could see linear scaling as the best-working scenario, but this can be harmful in introducing new values during the machine learning phase. The choice is yours to make… but normalization is definitely needed with the number of outliers your dataset contains.

Figure 3 shows that some genes correlate with each other. You assume that all the data have similar levels and distribution in correlation. You could strengthen that assumption by constructing a PCA plot based on the clustering methods of hierarchical cluster analysis based on absolute correlation. The labels of the plot are hard to read. This can be solved by adding some space between labels or removing the front end of a label's name (NP_). The x and y labels could be removed or renamed properly; they lack in this state.

Moving on to figures 4-6, we see a couple of density plots. Here is the expression of a specific gene under various scenarios displayed. Looking at the difference between all the three randomly selected records, we can state that normalization is a definite must. Maybe you can show more records to solidify a possible approach further to remove outliers. You could present these new records in a single graph using "grid.arrange" or another function from the grid library.

In this EDA, you specifically looked at the expressions of multiple genes. Although this is relevant for your machine learning project, it could be interesting to study some other attributes. For example, what the distribution of age is, or if there is a correlation between age and a certain expression value of a gene, you could potentially do this by performing a scatter or PCA plot. Underlining probable correlations can make your machine learning algorithm more trustworthy.

Your conclusion's formulation seems very good, but the dataset still needs some work before it's ready for machine learning. I hope my tips will help you to improve your EDA.