VA INTERNSHIP FINAL REPORT

JAY ALEXANDER TREVINO

Please note that many of my answers are adapted from my notes in the Jupyter Notebook, as I aimed to answer these questions in the notebook as well.

1. This data, like our real data, may be messy, incomplete, and/or sparsely documented. Please walk us through how you cleaned up this dataset. Do you see any interesting trends?

Initially, I noticed that the "Grade" column was not labeled as it should be. The description of the data states that the tumor grade is either 1-4 or unspecified. The value 9 could also be a typo of the value 1, but since I couldn't confirm this without seeing the original method of data collection nor consulting anyone immediately with domain expertise, I saw it best to drop this column.

The columns "T", "N", and "M" were of great interest to me and given the short time I had to complete this assessment, I approached these features by converting "T" and "N" into the proper dtypes as needed, notably making them numeric and dropping the letters present. Again, I felt that it would be sufficient to classify those points with succeeding letters as just the numeric values attached to them. For the "M" feature, I noticed that the data was sparse and also heavily skewed towards one of the feature's classes. Likewise, I calculated the survival rate of patients for cases where this feature was equal to just one or the other and saw no significant difference in the two, albeit one of the classes had a small sample size. I decided to drop this column.

For the "Stage" feature, I took the same approach that I did with the TNM features and only considered the numeric value so that I could convert it to a more manageable data type for the ML model. For the "Radiation" column, I converted this into a boolean 0/1 flag.

2. Tell us how you decided which features to use for your model. Are there any new features it might be productive to engineer from the current features?

I did add numerous columns for a variety of reasons:

First, I one-hot encoded the "Histology" feature since this feature is categorical and could be better represented in this way. I also did this knowing that the dataset was already small and took advantage of how fast the program could run because of this.

Next, I heavily considered the importance of the different genes present and the genomic dataset. I calculated the survival rates of having different numbers of mutated genes and found no significant difference in this analysis. However, I did notice that the survival rates of patients who had specific mutated genes did vary significantly and decided that I could best represent this by taking the genes present in at least 20 patients in the dataset, assigning them a unique column, and a flag to denote if a patient has this specific gene. Last, I found that there was a small difference in survival rate for patients who had a different number of mutations from the number of mutated genes (essentially at least one of the mutated genes a patient has will contain more than one mutation. I also assigned a column to whether this was true or not for a patient, and then proceeded to drop the columns pertaining to the number of mutations/mutated genes.

I also created a new column for whether or not a patient had cancer in both lungs using the "Primary.Site" feature. I made a quick judgement and assumed that cancer in both lungs would not lead to a promising outcome for the patient and added a flag for this, despite the small sample size for these cases.

3. Which algorithm did you use for the prediction and why?

I decided to use a random forest algorithm, specifically the *sklearn.RandomForestClassifier* because of the power of ensemble learning and bagging. I figured a decision tree-based classifier would allow me to tune the hyperparameters easily if needed since I have more experience with this algorithm than some others. I chose the "entropy" criterion because it might provide better results at the cost of runtime, however, runtime wasn't an issue since the dataset was small so I took advantage of this again. I also lowered the number of estimators from 100 to 10 to try and avoid any issues of overfitting since the data already only contained 190 data points.

4. How did you assess the predictive model's quality? Summarize your findings.

I only assessed through accuracy as the model was performing well and the rest of the analysis took up much of my time.

5. Next steps? What might you do with more time or access to additional data or expertise?

Should I have more time, I would definitely like to run an analysis on the correlation between variables. Especially with the TNM features, should there be some correlation between them, then it could be helpful to use those features to predict null values. I handled most cases of null values using the mean, but I can see better alternatives using smaller models and domain expertise to fill in these gaps. I also mostly ignored the "Primary.Site" feature as I mentioned above, but more time would allow me to calculate the survival rates and deduce if it's worth one-hot encoding this feature. Lastly, I would have liked to tune the hyperparameters using a validation set, but decided against it in the interest of time.

Thank you for this opportunity and I genuinely hope you enjoy this report and my notebook!
-Jay