# Methodology and Cross Validation Poster Information

In order to use the Random Forest Classification Model from the Sci-Kit Learn library to predict the vital statuses of patients, it was necessary to first process the data and remove irrelevant features. From the 1134 samples of colorectal adenocarcinomas, a subset of patients who had both liver metastasis and Level R1 actionability data was identified using a combination of selection brackets and .loc[] on the original dataframe containing the full list of samples. From there, taking a look at the dataset columns, it was decided which columns should be removed based on their relevancy towards impacting the vital status of a patient, the desired output of the mode. Next, since the Scikit-Learn is a library that only works with numerical types, it was pertinent to convert all non-numeric values into an appropriate numeric representation. In the case of ordinal data which described a logical progression, these values were directly converted into their numeric equivalent; alternatively, two columns that represented the answer to a binary question, such as: “was the tumor a left-sided tumor?” were merged into one column with 0 and 1 values. All other values with non-ordinal or binary data were one-hot encoded and then the whole dataset was put through a Random Forest Classification model. To determine the effectiveness of the model produced by the provided input, the average percentage accuracy of the theoretical output versus the true output was taken over the course of 10 iterations; to determine the most important features, the same process was performed and then afterwards the features were sorted in terms of their importance in descending order. Such information could provide valuable insight towards identifying relevant features that impact the survivability of patients.