For my capstone project I will be generating a predictive model for the diagnosis of breast cancer using a dataset consisting of features extracted from a mixed sample of malignant and benign tumor biopsies. The goal of this model is to increase the ease and confidence of triaging cancer patients so that the best use can be made of the available time and resources. This model must be strongest (ideally close to 100% recall) in identifying true positive (actual malignant) samples so that it does not make mistakes in these cases.

The data already consists of all numerical features (aside from the target) so the focus will be on building a strong and compact model to predict on new samples. All features are specific to nuclear attributes and so are directly relevant to the problem at hand, though more exploration will be done to evaluate whether there are any troublesome relationships. After processing the dataset for modeling, the bulk of the problem will be in model selection and tuning, and feature selection to ensure the best performance on unseen data.

One constraint for implementation is that the data are still fairly narrow (constrained only to an analysis of physical attributes of cell nuclei) so it is likely that this model will become part of a more extensive system of predictive tools for this use case. When presenting this solution, it would be helpful to frame it as a strong first step, rather than a standalone magic button for triaging. The stakeholders in this project would consist not only of doctors, but also of medical boards who are in charge of policy and funding, as well as the patients themselves, who should be well-informed on the mechanics behind this model and its performance statistics.

(The data are publicly available for download and contained within a single .csv file.)