# Capstone Project 1 Project Proposal

## Problem Definition

What is the problem you want to solve?

Predict if a patient has breast cancer based on the characteristics of cells derived from Fine Needle Aspiration procedure[[1]](#footnote-0).

## Stakeholders

Who is your client and why do they care about this problem? In other words, what will your client do or decide based on your analysis that they wouldn’t have done otherwise?

Health professionals are the clients. They will use the model prediction to determine the patient treatment plan.

## 

## Data Context

What data are you using? How will you acquire the data?

The dataset consists of 699 instances of cell characteristics from FNA procedure. It also includes whether the cells in each instance come from a malignant or benign tumor.

The dataset was aquired from UCI Machine learning repositor: [Link](http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29).

|  |  |
| --- | --- |
| Input Variables | Output Variable |
| 1. Clump Thickness 2. Uniformity of Cell Sizes 3. Uniformity of Cell Shape 4. Marginal Adhesion 5. Single Epithelial Cell Size 6. Bare Nuclei 7. Bland Chromatin 8. Normal Nucleoli 9. Mitoses | 1. Tumor Classification: Malignant or Benign |

## Plan for Solving the Problem

Briefly outline how you’ll solve this problem. Your approach may change later, but this is a good first step to get you thinking about a method and solution.

All the variables in this project are qualitative in nature. The author considered the following algorithms in evaluating which one to apply for this problem:

1. K Nearest Neighbours
2. Support Vector Machines
3. Linear Regression
4. Random Forests

Random forests were chosen based on their success to classification problems and ease of understanding. The Scikit learn routines also provide a measure of which input variables are important allowing for interpretation of the model.

Model will be built using 90% of the data. The remaining 10% will be used as test data.

To evaluate the model’s performance numerically, a confusion matrix will be built using test data.

|  |  |  |
| --- | --- | --- |
|  | Actual Benign | Actual Malignant |
| Predicted Benign |  |  |
| Predicted Malignant |  |  |

## Deliverables

What are your deliverables? Typically, this includes code, a paper, or a slide deck.

The deliverables include:

1. Code hosted on Github.
2. Slide Deck.

1. Fine Needle Aspiration is a diagnostic procedure in which needle is inserted into a body organ (e.g. breast) to collect a sample of cells. The cells are then examined for their characteristics to determine if they are malignant or benign. This procedure is a much safer and less invasive alternative to surgical biopsy. [↑](#footnote-ref-0)