Proposal

Breast cancer diagnosis prediction

Problem

Breast cancer is the second-ranked fatality cause in United State. However, breast cancer treatment is highly effective if it can diagnose at an early stage. A fine needle aspirate (FNA) is one of the quick diagnostic methods. However, it is not 100% accurate, and the study shows false negative 5-10% resulting late treatment procedure. Through this project, we would like to explore machine learning algorithms to find the most accurate model with a low false negative.

Target observer

Mainly doctor will be our observer for this project.

Data

The data of fine-needle aspiration can be found from:
LINK: https://www.kaggle.com/uciml/breast-cancer-wisconsin-data

Approach

We will approach this project by using python for coding. We will first transform the data into first cleaning the data and searching and replacing any null data. Then we will analyze the data set with aid from the seaborn such as plotting histogram and heatmap. The target data seems to have a binary outcome (malignant or benign), so we will explore a machine learning algorithm based on a binary classification algorithm. We will then compare all the models created to investigate the most accurate model. Also, we will use the shap library from python to investigate the most important feature which affects the prediction outcome.

The use

We are hoping this project could aid doctors to make a diagnostic decision if a patient has breast cancer.