Machine Learning Project

Carley Engleson
Cheyenne Martin
Ricci Sandoval
Ryan M

Key questions - Can we construct a machine learning model to predict whether breast cancer is benign or malignant based on multiple points of measurement data collected from histopathologic samples of cell nuclei?

Our dataset:
https://www.kaggle.com/uciml/breast-cancer-wisconsin-data

We will be retrieving and preprocessing data from the Breast Cancer Wisconsin (Diagnostic) Data Set, which uses key features of breast mass cell nuclei taken from fine needle aspiration biopsies. Our goal is to find the best machine learning model (which would have the highest testing data accuracy percentage) for this predictive analysis. The models we will be using are:

- Logistic Regression
- Support Vector Machines (SVM) before & after hypertuning the model
- Kernel SVM
- Random Forest Classifier

Here's how our models did by descending accuracy percentages & confusion matrix results:

1. Logistic Regression → **98.6%, 0 FP, 2 FN** (overall best model)

2. Hypertuned Support Vector Machines (SVM) → **97.9%, 5 FP, 1 FN** (had the highest overprediction of malignancy, but had very low underprediction of malignancy)

3. Random Forest Classifier → **97.2%, 1 FP, 3 FN** (tended to overall underpredict malignancy)

4. Kernel SVM → **96.5%, 3 FP, 2 FN** (better on avoiding underpredicting malignancy)

5. Support Vector Machines (SVM) → **95.8%, 4 FP, 2 FN** (2nd highest overprediction of malignancy, but still had low underprediction of malignancy)