Capstone Project 1: Milestone Report

Mehmet Erdoğdu

If breast cancer is left untreated, the cancer spreads out to other parts of the body if it is a malignant cell growth. Benign cells are usually localized and do not spread to other parts. I will predict if a given cancer cell is benign or malignant to be able to treat the cancer cells in a timely manner. The clients for this project would be hospitals and medical institutions. They care about this problem because using a highly accurate prediction model can reduce lives lost due to cancer by taking necessary preventive actions on malignant cancer cells.

The Dataset

I will be using the University of California, Irvine Machine Learning repository Breast cancer diagnostic data set. I acquired the data set through Kaggle website, which can be found at this link:https://www.kaggle.com/uciml/breast-cancer-wisconsin-data

The dataset is pretty clean, straightforward but I still did an exploratory data analysis for the feature selection. The data set has 33 columns with 569 rows. There are ten computed real-valued features for each cell nucleus in the data set: radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, fractal dimensions. The mean, standard error and "worst" or largest (mean of the three largest values) of these features are also computed resulting in 30 features. There are a total of 357 Benign and 212 malignant points in the data set.

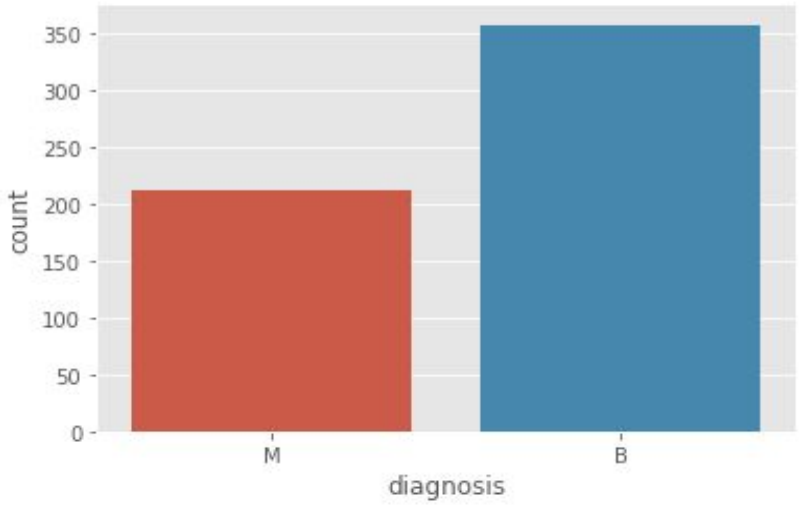

Figure 1 Benign and malignant groups

Reducing Features

As each feature has 3 corresponding columns - worst, mean, and standard error, I thought it prudent to mitigate multicollinearity which may result. As such, I used logistic regression to select the best predictor in each group of features between worst, mean, and standard error and found the following to be the best predictors:radius worst, texture worst, perimeter worst, area worst, smoothness worst, compactness mean, concavity mean, concave points worst, symmetry worst, fractal dimension worst. Since I don't want to assume that there were no good features in each group other than the "best" one, I have also calculated the correlation between the best feature and the other two features in the same group to confirm that multicollinearity is an issue. If the correlation were smaller than .5 between the two features I would also add it to my best features list. The final list consists of 13 features.

Exploratory Data Analysis

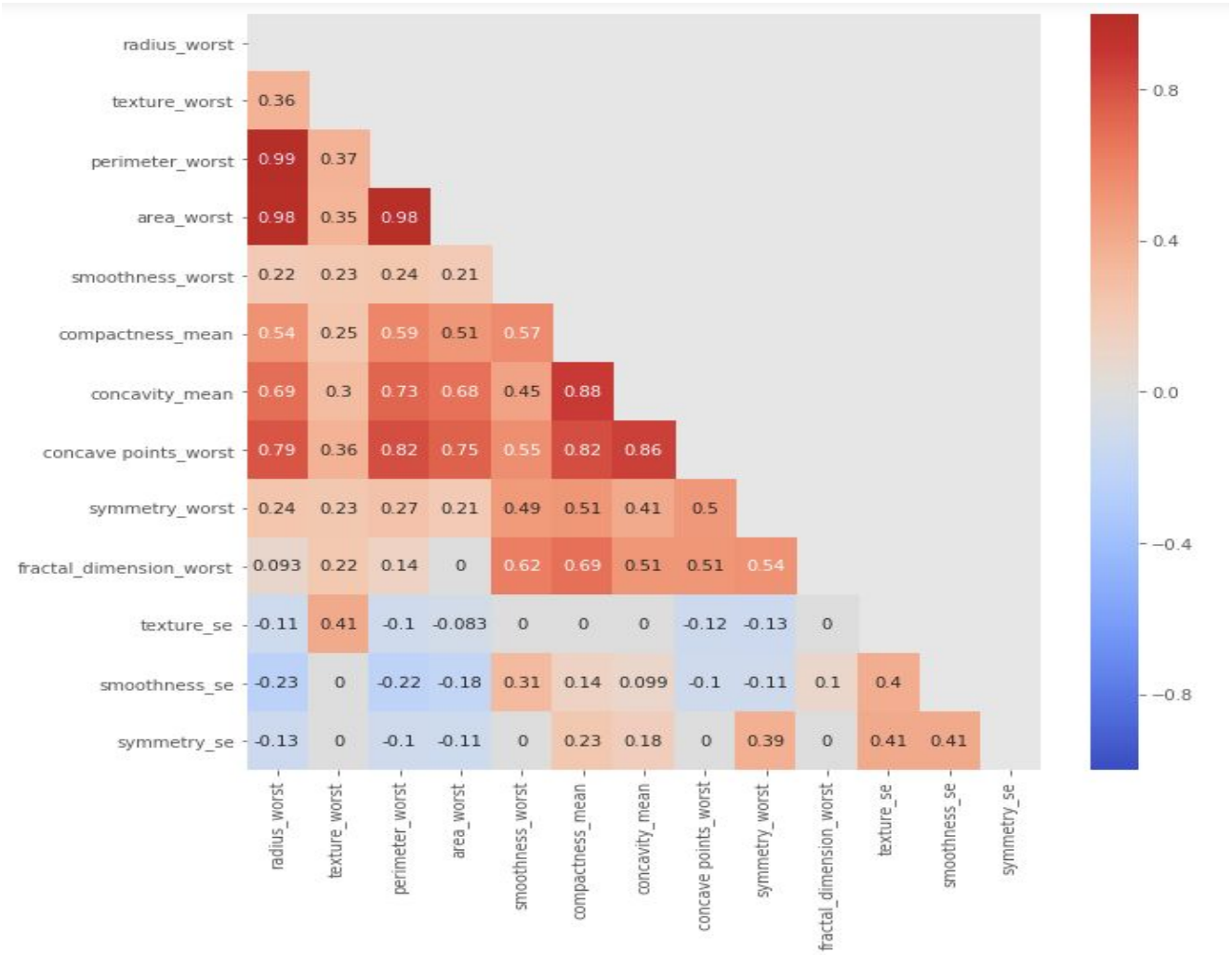

Figure 2 Correlation matrix of the best features

Above, you can see the correlation matrix of the best features. I have grayed out non-significant correlations between each two features group combinations.

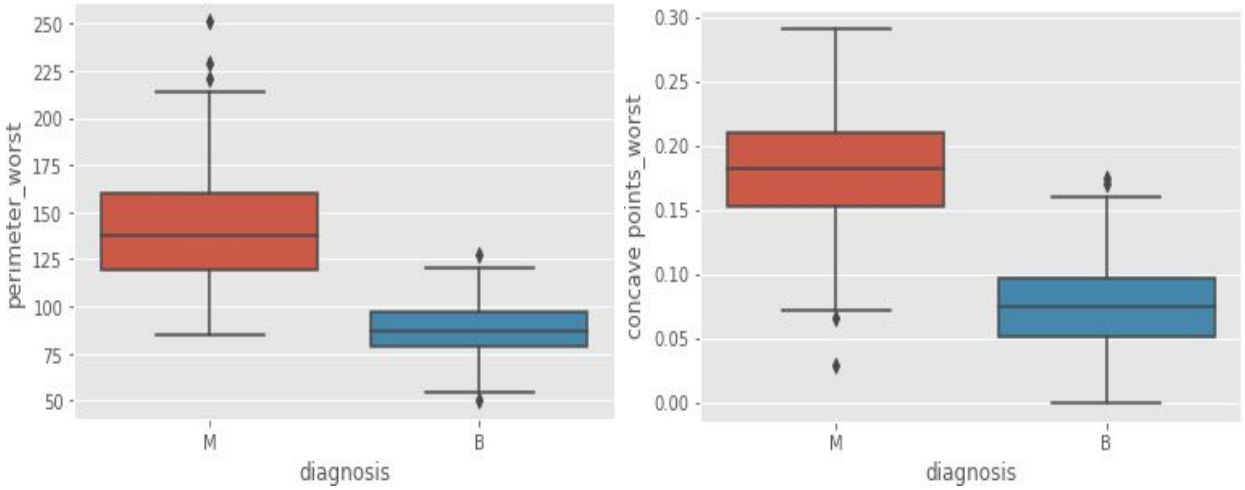

Figure 3 Perimeter and Concave points box plots

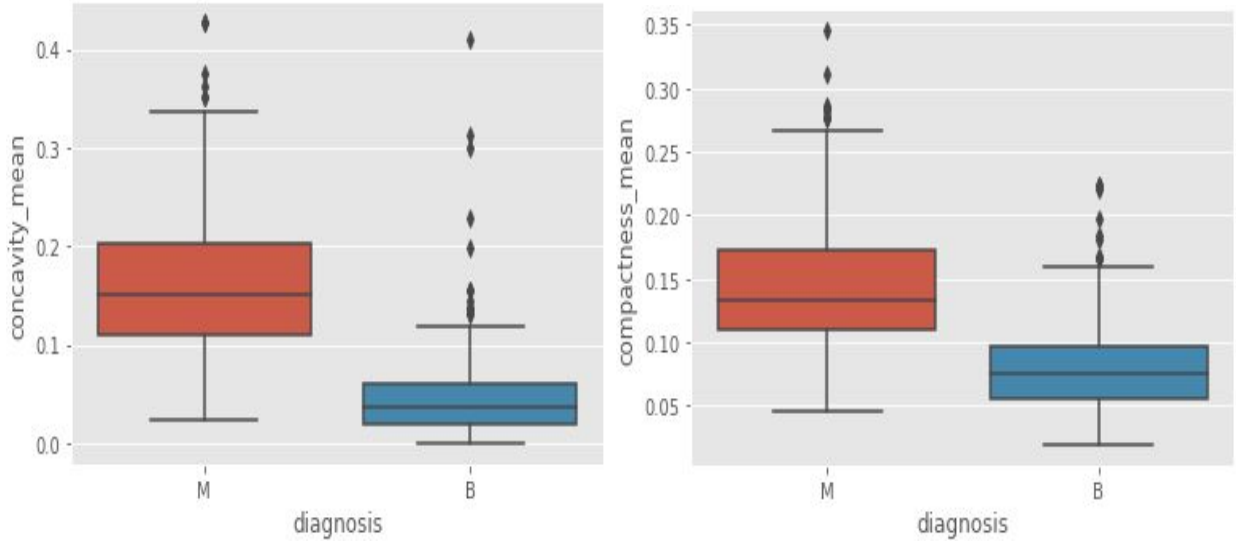

Figure 4 Concavity and compactness box plots

As the boxplots above show, Malignant and Benign cells tend to have different characteristics which makes it seem promising that they can be distinguished with a predictive model. In particular, concavity, perimeter, compactness, and number of concave points are all larger in Malignant cells than in Benign cells. Statistical testing using a t-test confirms that this relationship is statistically significant across all of these features with $p < .05$.