Figure One: Features and Their Distribution for Malignant and Benign Cells

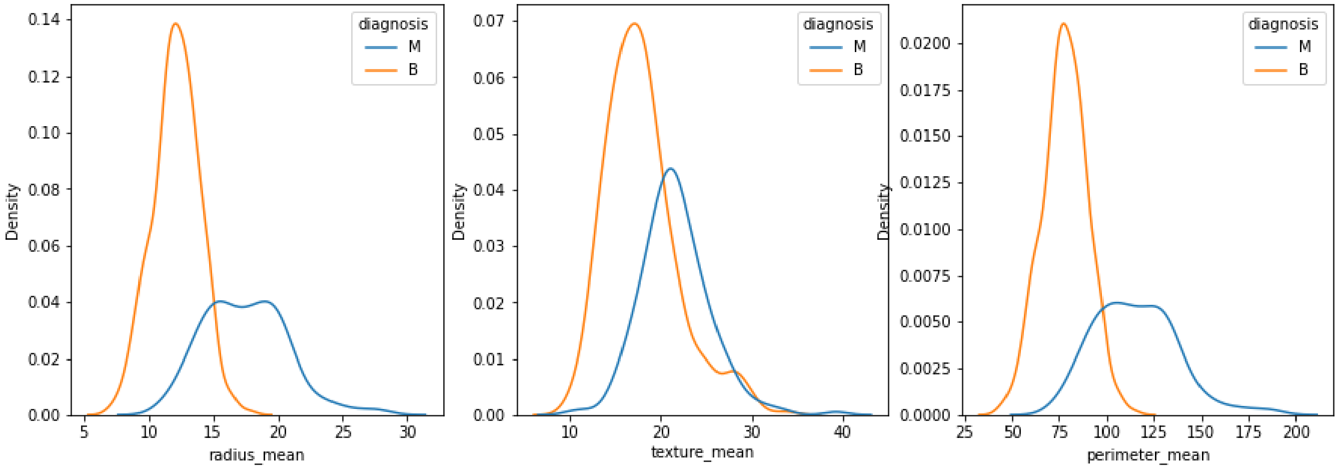


Figure Two: Principal Components 1 and 2 and their Diagnosis

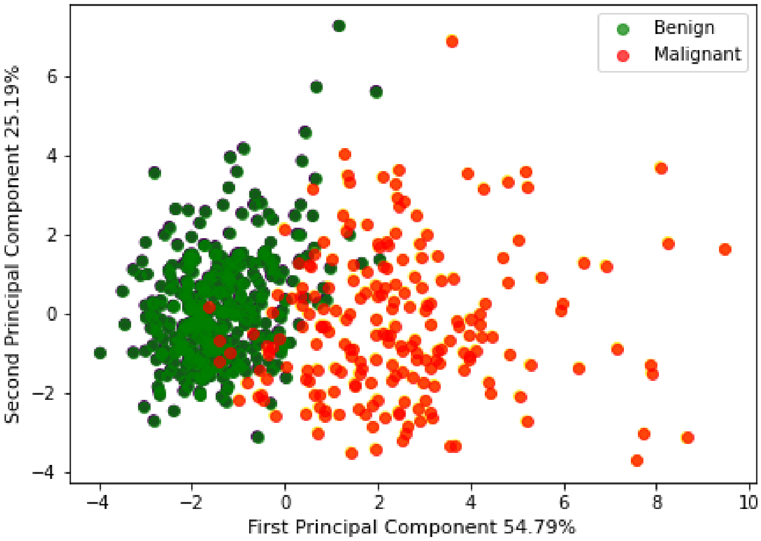
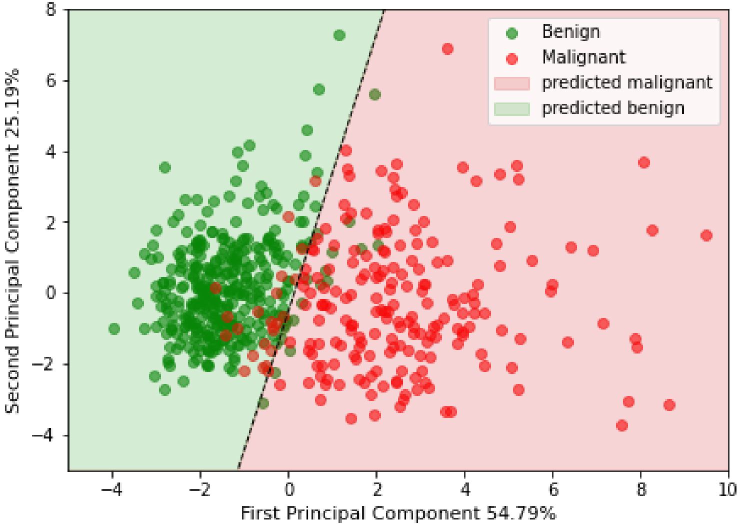


Figure Three: Principal Components/ Predicted vs. Actual Diagnosis



Breast Cancer Predictions
Amanda Burger

The most common type of cancer in the United States is breast cancer, ranging over 250 thousand new cases every year. Although testing and treatment is becoming much easier and more accurate, there are features of the cells within the tumors that can help us determine whether it is benign or malignant. The goal of my project today is to see if these features can accurately predict diagnosis. Identifying these features can allow doctors to more easily detect whether a tumor is malignant.

The data is from Dr. William H. Wolberg that works in the General Surgery Dept at the University of Wisconsin, Clinical Sciences Center. The data is web scaped from https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29 The features in the dataset were computed from a digitized image of a breast tumor, taking the values from each nuclei and reporting the mean for each of the tumors. The dataset analyzed 596 tumors with 12 columns that include the ID, diagnosis, and 10 features like radius mean, texture mean, smoothness mean, etc. Every field is populated.

For analysis we select all the features and evaluate their effects on diagnosis. These features include: (1) radius mean, (2) texture mean, (3) perimeter mean, (4) area mean, (5) smoothness mean, (6) compactness mean, (7) concave points mean, (8) symmetry mean, (9) concavity mean, and (10) fractal dimension mean.

For **Figure One** we can see how each of features were distributed for both types of diagnosis. The given graphs have the features value on the x axis and percent of data at that value on the y axis. If we look at the radius mean graph for example, we can see that lower values of this feature are more likely to be benign than upper values. We also see that malignant tumors have a greater range of radius values than benign ones. This is a trend for all of the features too though, not just these 3. In general benign tumors have features with lower values and less of a range, while malignant tumors have higher values with a wider range. So in conclusion, for all the given features, higher values indicate the tumor to be more likely malignant.

Figure Two is a PCA plot that plots the first and second components of the PCA transformed data. As the plot shows, both the components together account for about 80% of the variance. Transforming the data from 10 variables to two dimensions, allows us to see the relationship between the features and diagnosis. As you can see from the figure, the 10 variables selected are a good indication of whether the tumors are benign or malignant. The plots' clear divide in the diagnosis along with high variance ratio shows us that the given features are a great predictor. The plot shows that lower values of the 1st component predict benign tumors with higher values indicating malignant tumors. This is similar to the density plots in figure one. Both plots also show that malignant tumors have a larger range of values than benign ones. Overall, the PCA plot tells us that by converting the 10 feature data into two dimensions we can see a clear connection/relationship between the given features and tumor diagnosis.

Figure Three calculates the relationship between the two PCA components above in a logistic regression. We use a pipeline to connect StandardScaler, PCA, and LogisticRegression to create a model that predicts whether the tumor will be malignant or benign. After creating a random 25/75% train-test split and training our data we can compare the actual to predicted values. The shading and scatter allows us to compare the model to actual values. As you can see on the graph the divide seems fairly accurate with a precision of 87% and a recall of 94%. The high recall tells us that the model is accurate, especially when detecting positive (malignant tumors).

In conclusion we have discovered that the 10 given features are clear indicators in predicting whether a tumor is benign or malignant. We learned that higher values with a wider range relate to malignant tumors while lower values with less variance relate to benign tumors. Because of this we were able to create an accurate model to predict a tumor's diagnosis. Overall these indicators and this model could be an important tool in detecting breast cancer early in patients around the world.