Analyzing the Breast Cancer Wisconsin (Diagnostic) Data Set

Erik Halasz Mark Frankli

December 2024

1 Introduction

The goal of our project was to analyze a dataset containing information about real-world patients who either have benign or malignant breast cancer. This is a standard binary classification problem, and since the dataset turned out to be fairly good quality, we also explored further questions such as feature importance and model versatility. The data contains a wide range of features, including variables such as the radius and perimeter of the nucleus, in addition to properties about its shape and symmetry.

2 Exploratory Data Analysis

Firstly, a little background about the dataset: features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass, which is a standard procedure in such cases. For every feature, three records are created: one with its mean, one with its standard error, and one with the average of the three largest values (referred to as "worst"). Then, we started our investigation with a standard exploratory data analysis on the dataset. After this, we concluded that it does not contain any missing values, all the features are floats (except the target variable that we $0-1$ encoded numerically), is slightly imbalanced (357 benign, 212 malignant) and most features follow a normal distribution. The rest seemed to be positively skewed, so we applied a log-transform on those features.

Next, we applied a correlation matrix to the dataset to see which features may be important. It can be observed that the dataset is highly correlated and there are only a handful of features that actually influence the classification outcome: *perimeter, radius, area, texture, symmetry* and *concavity*. In general, the larger the cell is (which is measured by *perimeter, radius* and *area*), and perhaps more deformed, the more probable it is that the patient has malignant breast cancer since tumors tend to grow and invade surrounding tissue.

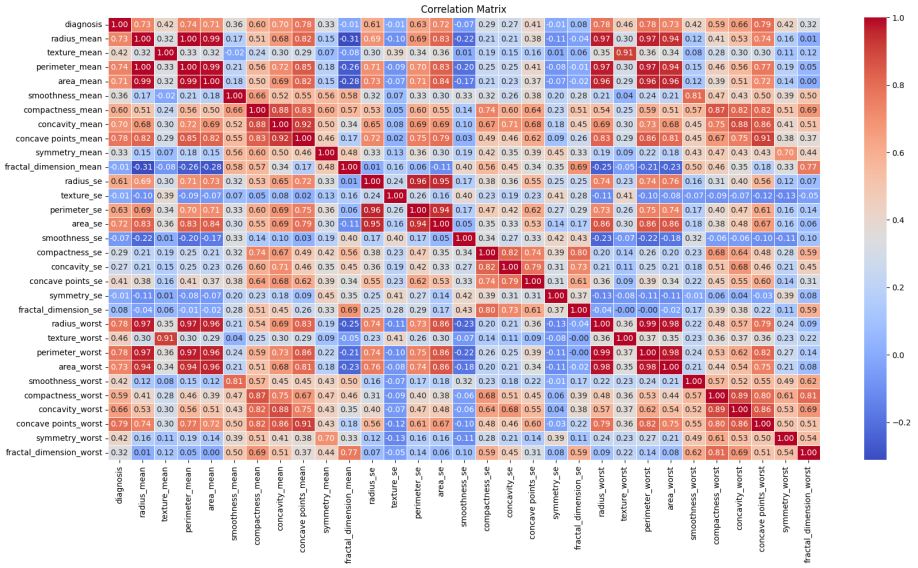

Figure 1: The correlation matrix of all the features. We can see that the data is highly correlated partly due to the fact that every feature is recorded three times.

3 Model Selection & Training

During the EDA, we observed that the data is well-separated, so a simple and natural candidate for our first model was KNN. Surprisingly, it achieved a approx. 95% accuracy, which is already impressive. Next, we applied three additional models: Neural Networks, Logistic Regression, and Random Forest. They scored similarly well, with Logistic Regression achieving the highest score of 99%.

In general, we used F1-score for our measure, an 80%-20% train-test split, stratified cross-validation, and grid search for hyperparameter optimization (where applicable).

4 Further Questions

Since the dataset seemed to be rather sterile and artificial, and the models also achieved a fairly high score, we also explored a few extra questions.

First, we wanted to know which features were the most important to test our hypothesis from the EDA. For this, we used two of our previous models, namely Logistic Regression and Random Forest, since they both associate weights/coefficients to different features during the training phase.

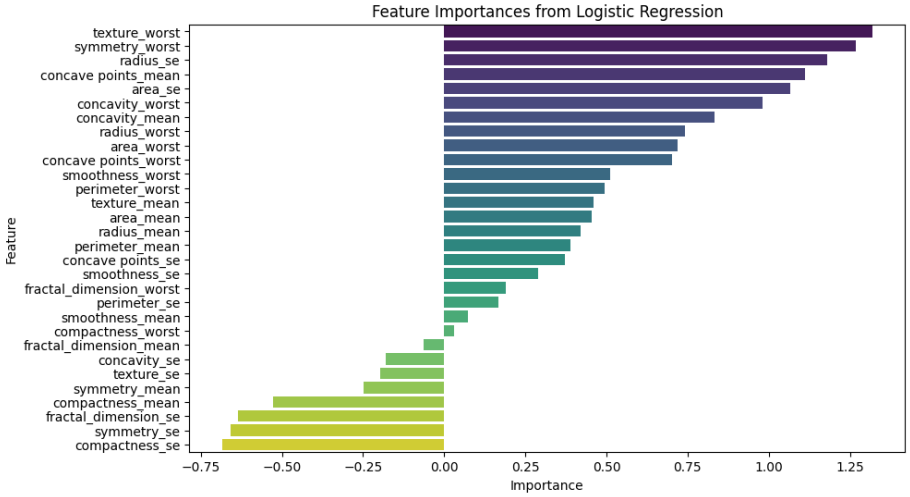

Figure 2: Feature importance extracted from Logistic Regression. We can see that the features that correlate the most to the target variable are the same in our hypothesis.

Next, we tested the robustness of Logistic Regression, Random Forest and Neural Networks by intentionally deteriorating the quality of the dataset. We added 5% of uniformly generated random noise to each feature, proportional to its range, only kept one feature per type (we arbitrarily chose the "mean") and replaced 30% of the existing data with NaN values. Also, we randomly selected a few features and dropped them. After this, we tested every model on this new dataset and found that their performance is almost the same as before (still achieving 90%+ accuracy), but if at least one of the important columns are dropped the F1-score falls below the 80% threshold.

5 Conclusion

At the end we make the following remarks:

- The dataset is outstandingly good: even simple models such as KNN achieves a 95%+ accuracy, while more complex models are close to 100%.
- There are only a handful of relevant features that actually correspond to the classification: radius, area, perimeter and concavity. This also aligns well with the biological interpretation since the larger the nucleus is the more probable it is that the patient has breast cancer (since malignant cells tend to invade their surroundings).
- Even after deliberately deteriorating the quality of the dataset the tested models still achieved 90%+ accuracy (with F1-score).