PROJECT PROPOSAL WITH EDA

BREAST CANCER WISCONSIN PREDICTION

Ameya Jamgade | BANA 8083 | May 26 2018

1. PROJECT DESCRIPTION

1.1 Introduction

Breast cancer is a malignant cell growth in the breast. If left untreated, the cancer spreads to other areas of the body. Excluding skin cancer, breast cancer is the most common type of cancer in women in the United States, accounting for one of every three cancer diagnoses. Breast cancer ranks second among cancer deaths in women.

This project aims at analyzing data of women residing in the state of Wisconsin, USA for prediction whether the case of breast cancer is malignant or benign.

1.2 Goal Statement

The project includes application of several data mining and machine learning techniques to classify whether the tumor mass is benign or malignant in women residing in the state of Wisconsin, USA. This will help in understanding the important underlaying importance of attributes thereby helping in predicting the stage of breast cancer depending on the values of these attributes. Through the understanding of nature of attributes in cancer prediction and prediction, healthcare community can take perform additional research corresponding to these attributes to help prevent pervasion of breast cancer into the population of USA.

1.3 Assumption and Scope

- The project assumes that the dataset collected is representative of the entire women population of Wisconsin, USA.
- The data has been collected accurately
- No errors have been committed while entering the collected data

The scope of the project is confined only to prediction of breast cancer to be malignant or benign for the women of Wisconsin only. This project will not include any conclusions that can be made whatsoever for the remaining women population of USA. The project will not go in depths of the reason why some of the attributes are more important that other attributes in prediction of breast cancer cases, as this would require considerable domain expertise on biomedical sciences.

2. PROPOSED SOLUTION OVERVIEW

2.1 Data Cleaning and Preparation

This stage would include identifying if the data set contains any missing values or bad data. I would the convert the diagnosis (Y variable) into appropriate format (currently it is classified as M and B for malignant and benign respectively)

2.2 Exploratory Data analysis

The EDA process will help in understanding the nature of the dataset and also help in identification of potential outliers or correlated variables.

2.3 Modelling

Since the dataset contains around 32 different attributes, I would try to incorporate principal component analysis to identify the important attributes.

For modelling, I am planning to use K-nearest neighbor, Random forest and Support Vector Machine algorithms for classification of the diagnosis Y-variable in the data as malignant or benign. Depending on the prediction accuracy and time constraint I might try to include Gradient Boosting Model technique in my analysis.

3. MEASURE PHASE

3.1 Dataset and data description

The data for this project is obtained from UCI Machine Learning repository. The link for the data can be found at:- https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

Features in the dataset are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. Separating plane described above was obtained using Multisurface Method-Tree (MSM-T) [K. P. Bennett, "Decision Tree Construction Via Linear Programming."

Proceedings of the 4th Midwest Artificial Intelligence and Cognitive Science Society, pp. 97-101, 1992], a classification method which uses linear programming to construct a decision tree. Relevant features were selected using an exhaustive search in the space of 1-4 features and 1-3 separating planes.

The dataset contains information of 569 women across 32 different attributes.

1) ID number
   2) Diagnosis (M = malignant, B = benign)

2) Ten real-valued features are computed for each cell nucleus:
   a) radius (mean of distances from center to points on the perimeter)
   b) texture (standard deviation of gray-scale values)
   c) perimeter
   d) area
   e) smoothness (local variation in radius lengths)
   f) compactness (perimeter^2 / area - 1.0)
   g) concavity (severity of concave portions of the contour)
   h) concave points (number of concave portions of the contour)
   i) symmetry
   j) fractal dimension ("coastline approximation" - 1)

The variables are divided into three parts first is Mean (3-13), Stranded Error (13-23) and Worst (23-32) and each contain 10 parameters (radius, texture, area, perimeter, smoothness, compactness, concavity, concave points, symmetry and fractal dimension). Mean is the means of the all cells, standard Error of all cell and worst means the worst cell

3.2 Data Cleaning

- The dataset contained one columns in which all the values were missing. This column was removed from the dataset. Apart from this the data didn’t contain any missing values.
- The dataset contained 0 duplicate values
- 62% are benign and 37% are malignant cancer cases in the dataset.

3.3 Exploratory Data Analysis

3.3.1 Breakup percentage of benign and malignant cancer cases

The dataset contains a class imbalance between the malignant and benign cases. Out of the 569 cases in the dataset, 62% cases are benign cases of cancer whereas 37% cases are malignant.

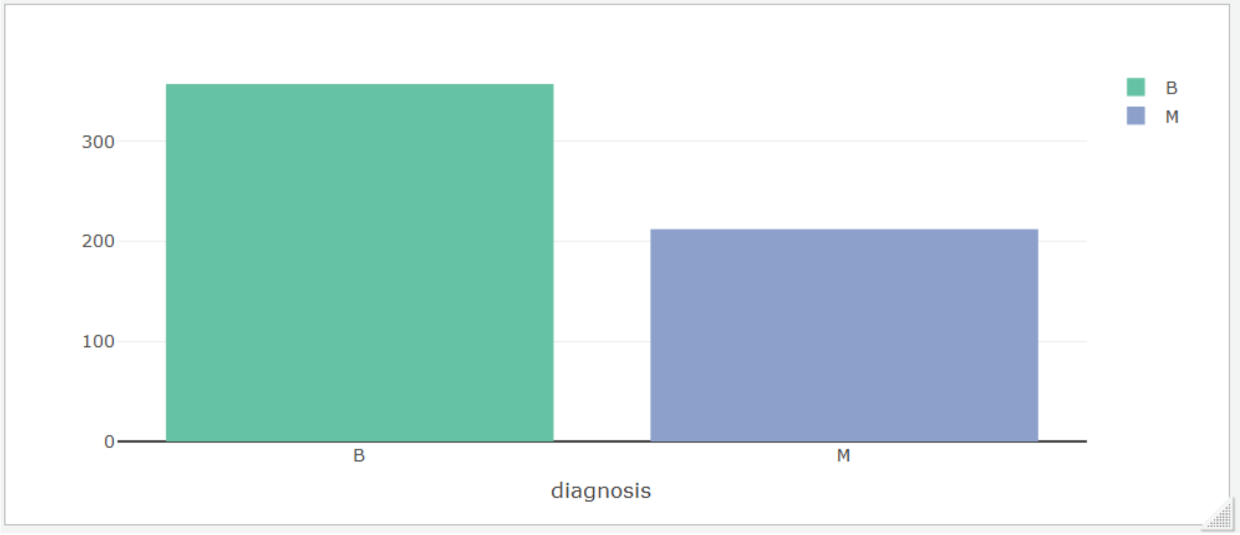

Figure 1: Breakup of benign and malignant cases

3.3.2 Correlation Plot

The first step in exploratory data analysis step is to identify if at all there is any correlation between any of the 32 variables. By using Pearson's correlation, the below plot was created.

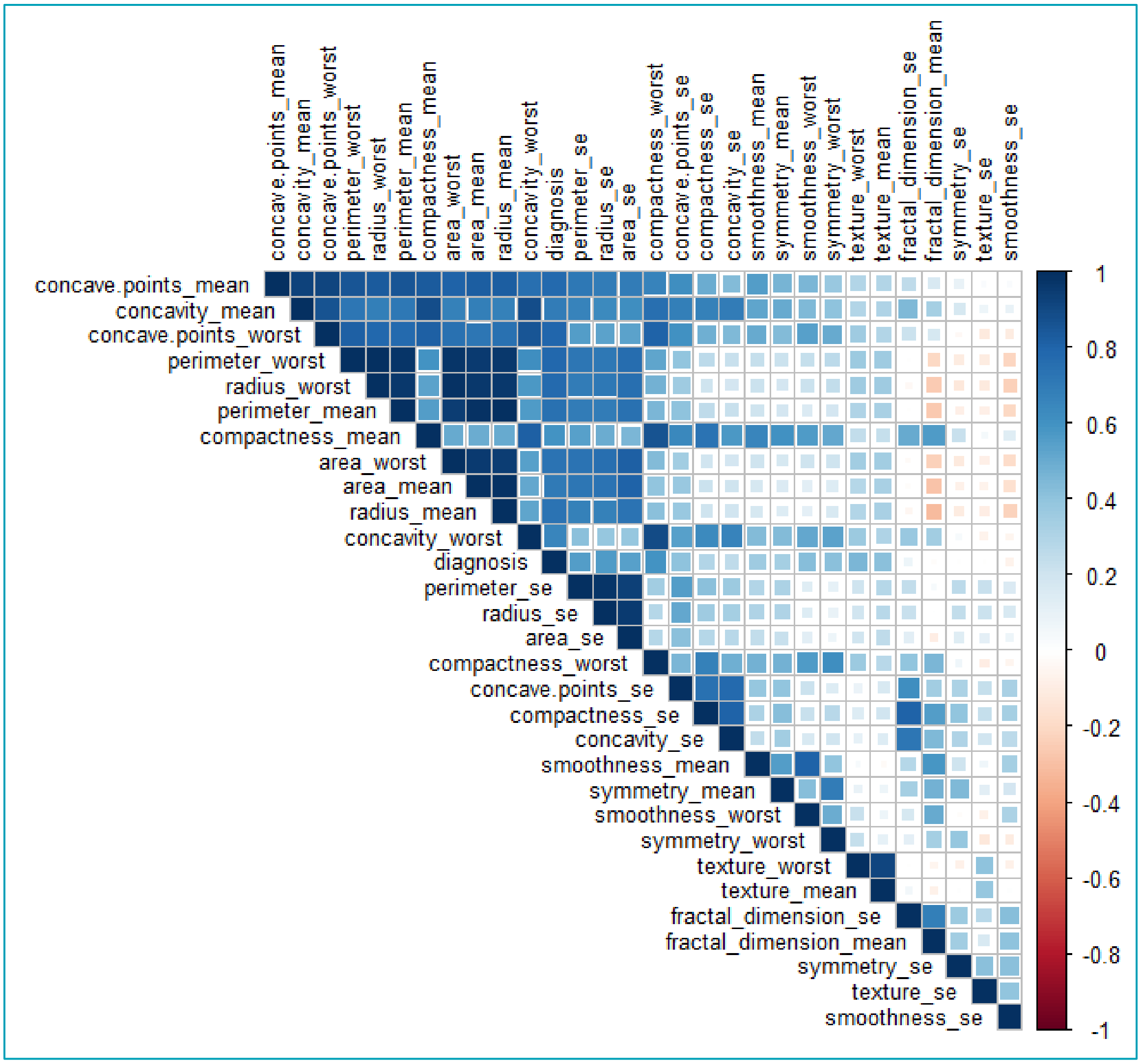

Figure 2: Correlation plot

We can see that the highest correlations are between

1. Perimeter_mean and radius_worst
2. Area_worst and raidus_worst
3. Perimeter_worst and radius_worst
4. Texture_mean and texture_worst

The next step in the analysis to to visualize some of these correlations as scatterplots in order the better understand the relationship between these variables mentioned above.

As can be seen from figure 3 below, the correlation between the two variable is very high at 0.969

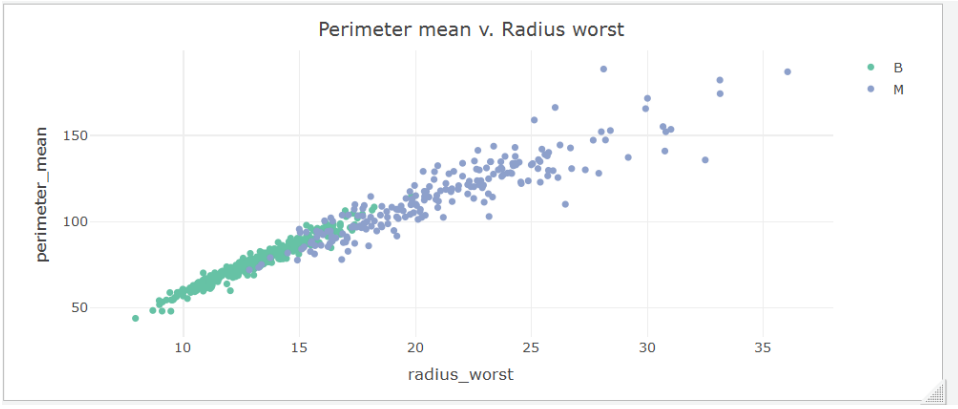

Figure 3: Correlation plot of perimeter mean and radius worst

As can be seen from figure 4 below, the correlation between the two variable is very high at 0.984. This is quite high and sufficient care must be taken while performing model building while using these two variables together.

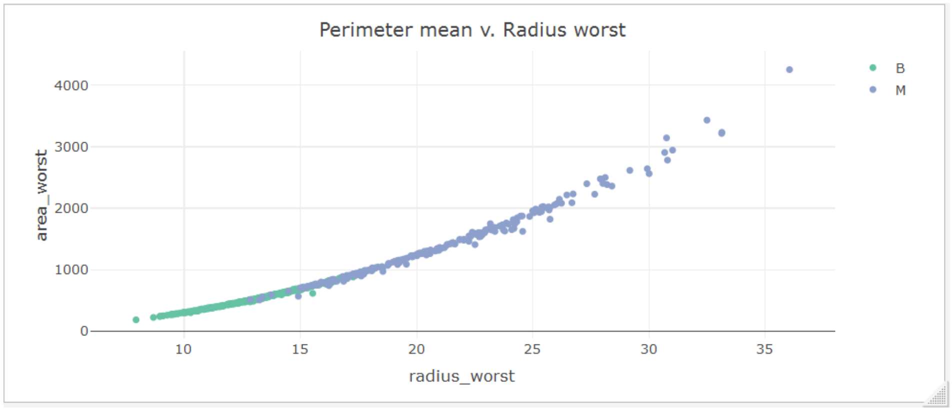

Figure 4: Correlation plot of area worst and radius worst

As can be seen from figure 5 below, the correlation between the two variable is very high at 0.993. This is quite high and sufficient care must be taken while performing model building while using these two variables together.

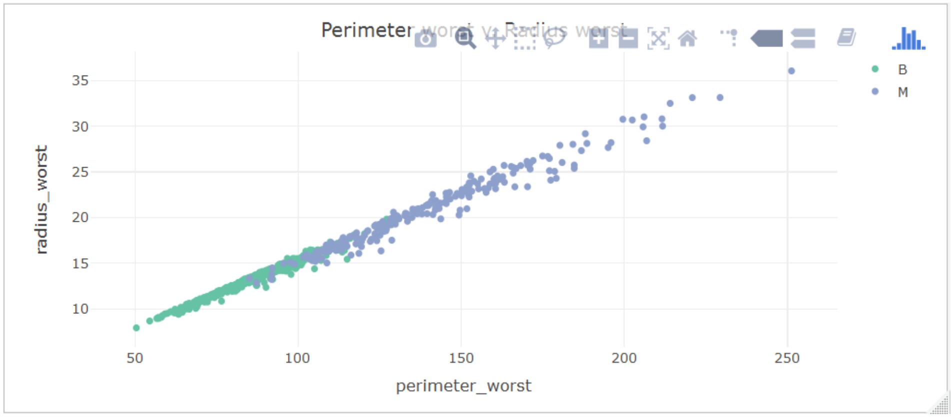

Figure 5: Correlation plot of radius worst and perimeter worst

As can be seen from figure 6 below, the correlation between the two variable is very high at 0.91.

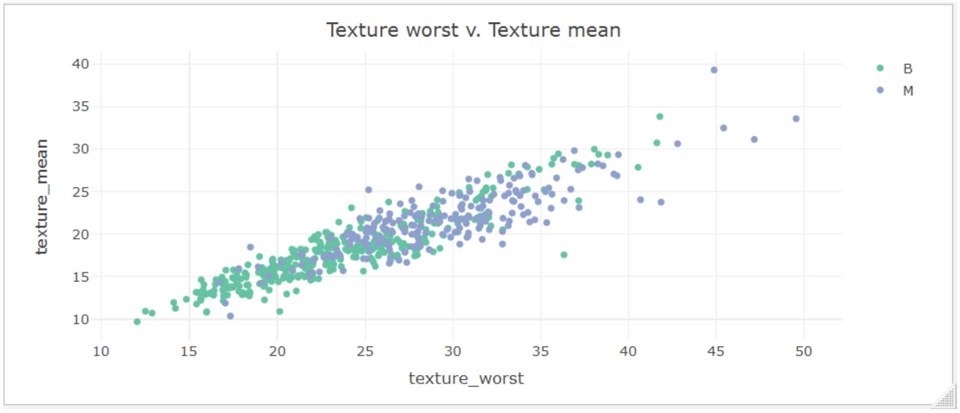

Figure 6: Correlation plot of texture mean and texture worst

3.3.3 Box Plot

In order to dig deeper into the dataset, it is necessary to see if there are any outliers in the data. For this purpose, I plotted boxplots of all the columns in the dataset. As can be seen from figure 7, area_mean, area_se and area_worst attributes have outliers present.

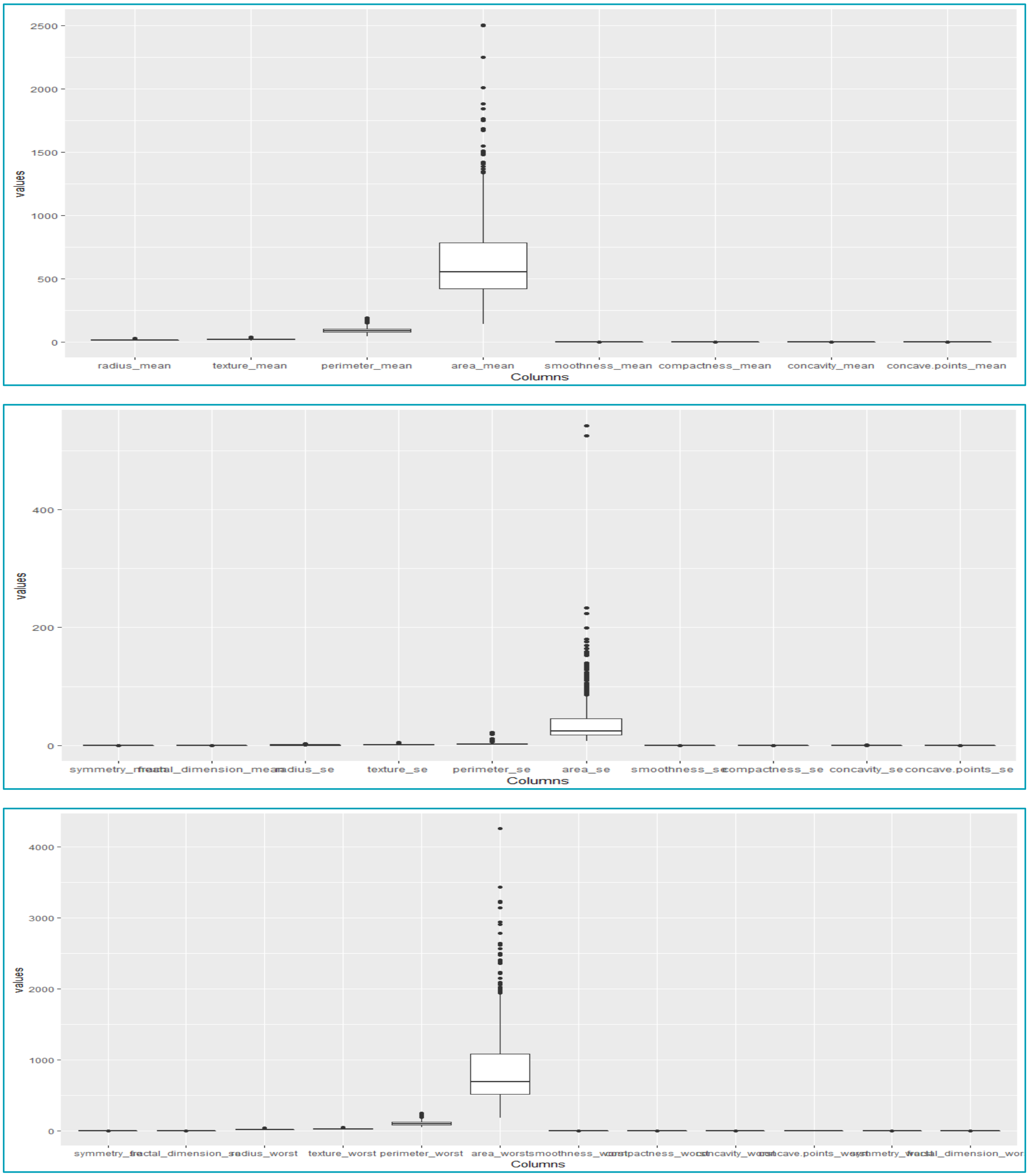

Figure 7: Boxplots of all columns in the dataset

Finally, in order to better understand the distribution of the data for each of the benign and malignant class, it becomes essential to plot boxplots for all the columns separated by benign and malignant classes. Figure 8 shows this. We can see that almost all of the attributes have outliers present in them when we break them into malignant and benign classes.

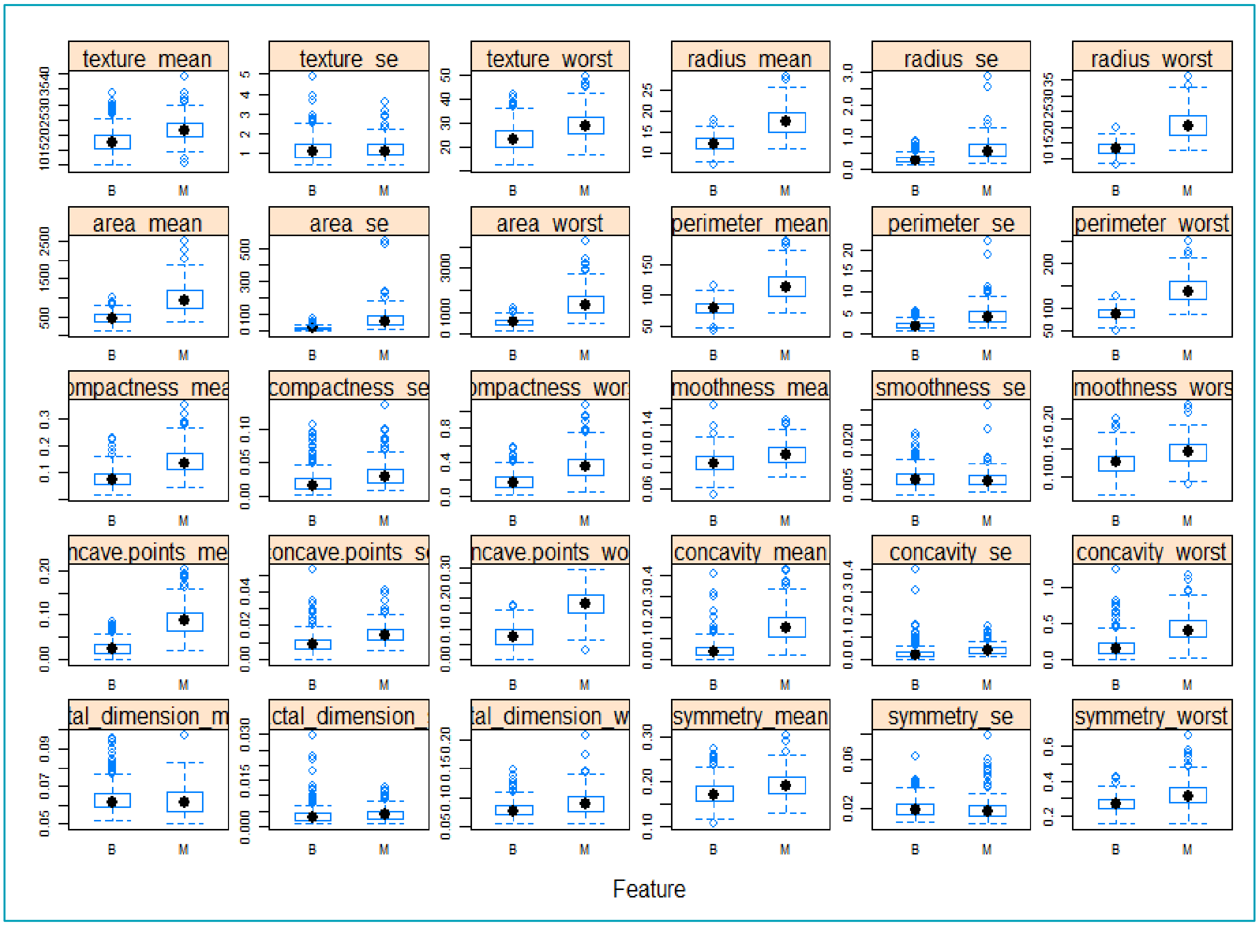

Figure 8: Boxplots of all columns broken down by class