Intro_logistic

Anyu Zhu

3/24/2022

Introduction

Objective

The goal of the project is to build a predictive model based on logistic regression to facilitate cancer diagnosis.

Dataset

The dataset ’breast-cancer’we used contains 569 rows and 32 columns. The variable `diagnosis` identifies if the image is coming from cancer tissue or benign. We labeled `malignant` as 1 and `benign` as 0. In total there are 212 malignant cases and 357 benign cases. There are 30 variables corresponding to mean, standard deviation and the largest values (points on the tails) of the distributions of 10 features: radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension.

The distribution of mean value of the 10 features among malignant cases and benign cases is shown in the plot below (Figure 1).

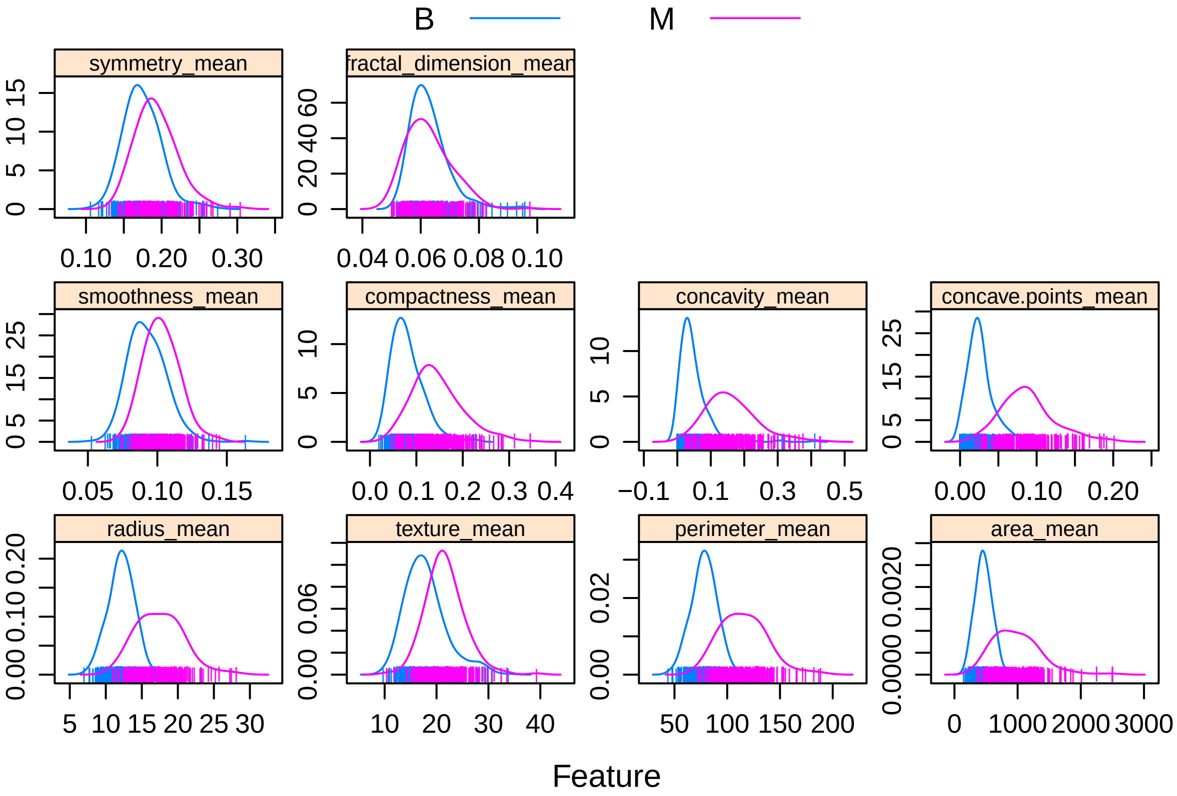
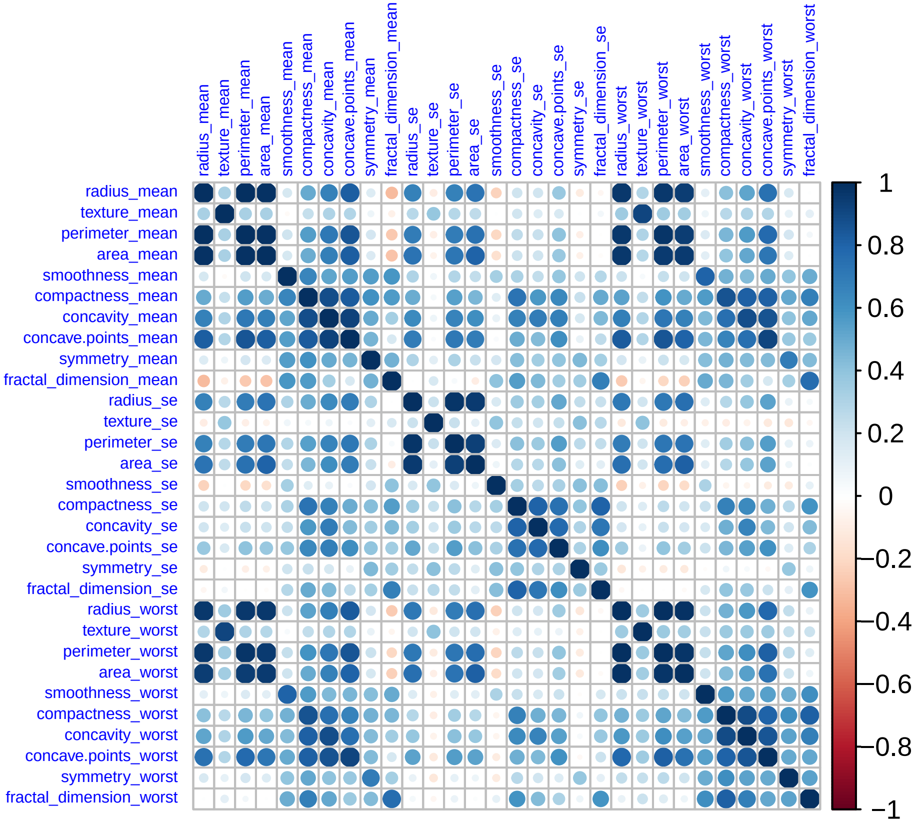


Figure 2 displays the correlation between variables. We can see there exists multicollinearity in the dataset. (?? Put this into discussion ??)

Figure 2: Correlation Plot



Method

Logistic Model