FEDERICO RUEDA

LUCAS WHITMIRE

EDGARDO PANZA

# INTRODUCTION

For this project we choose a dataset related to breast cancer detection. Lung and breast cancers were the most common cancers worldwide, each contributing 12.3% of the total number of new cases diagnosed in 2018.

In 2019 in the US, an estimated 268,600 new cases of invasive breast cancer were diagnosed among women and approx. 2,670 cases were diagnosed in men.

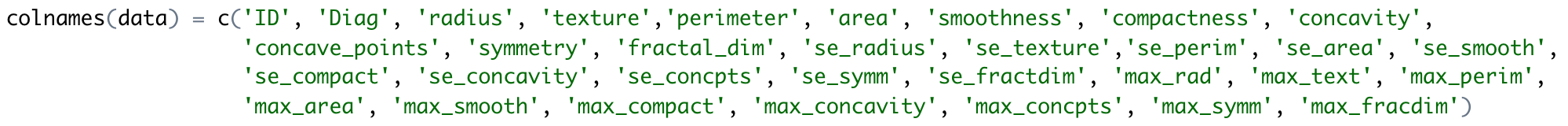
# OBJECTIVE

Apply in the chosen dataset concepts learned in the class of machine learning II and create the best model.

# DATASET

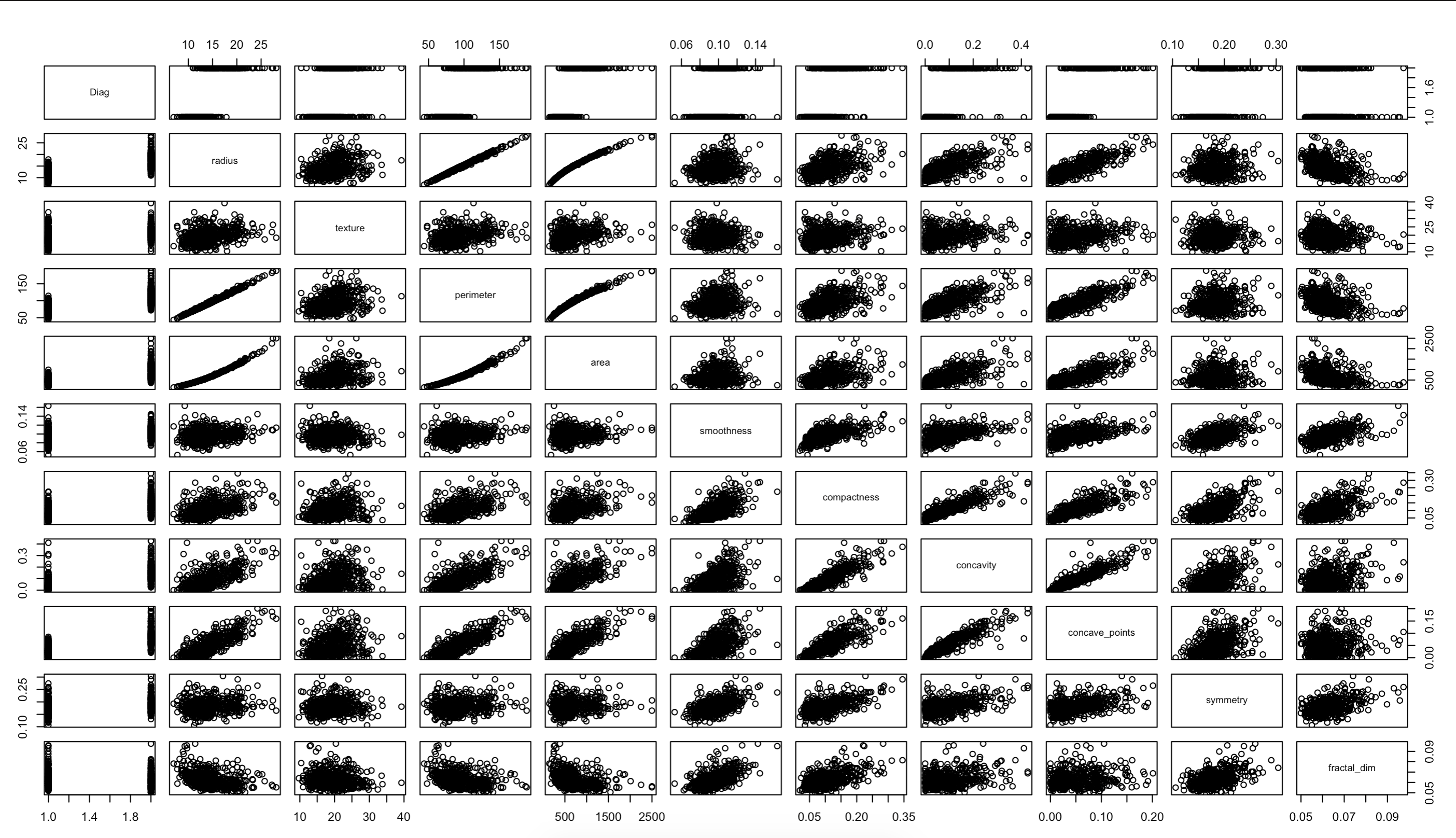
The dataset consists of 569 observations and 31 variables.

The variables are:

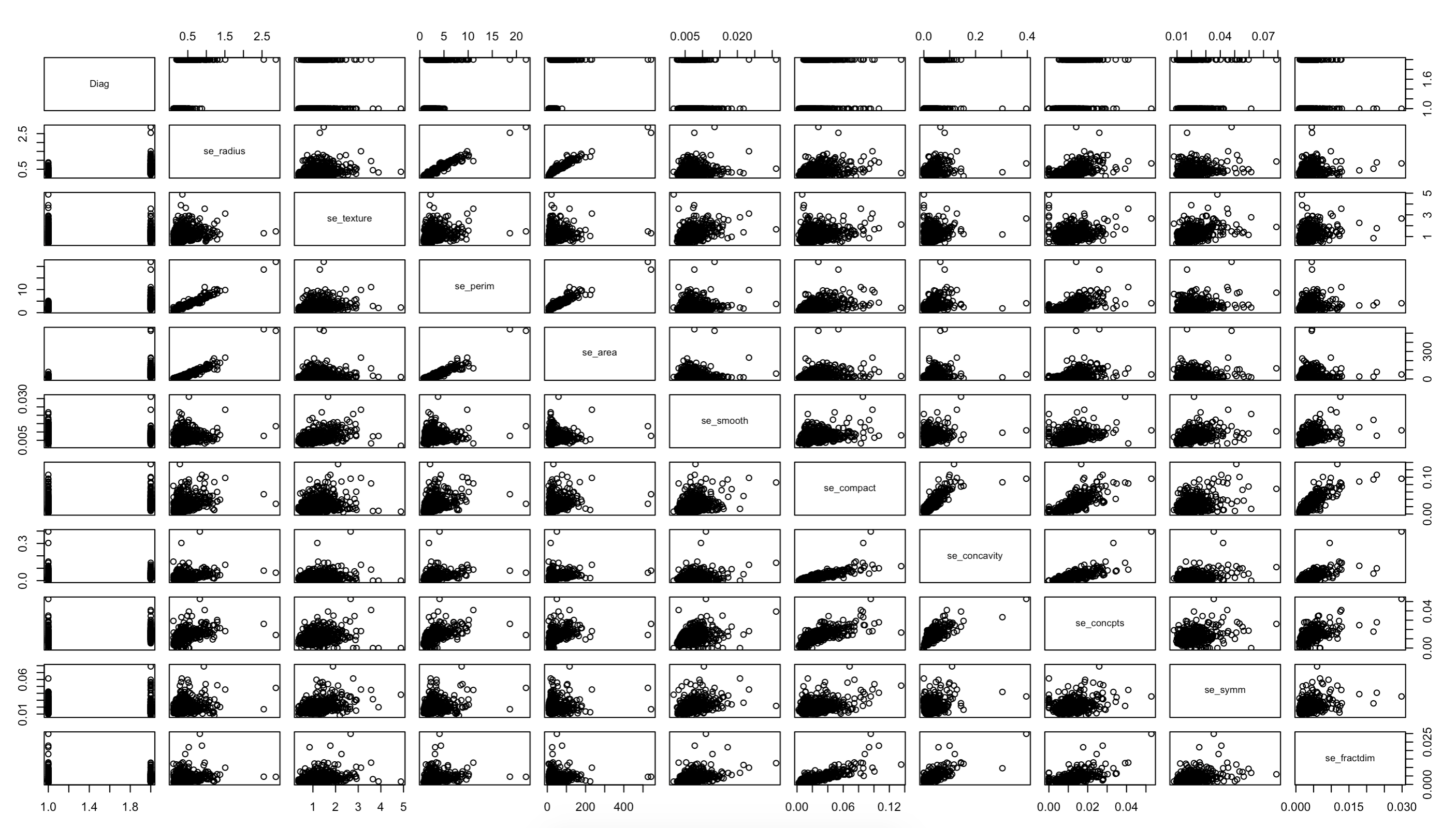


# EDA

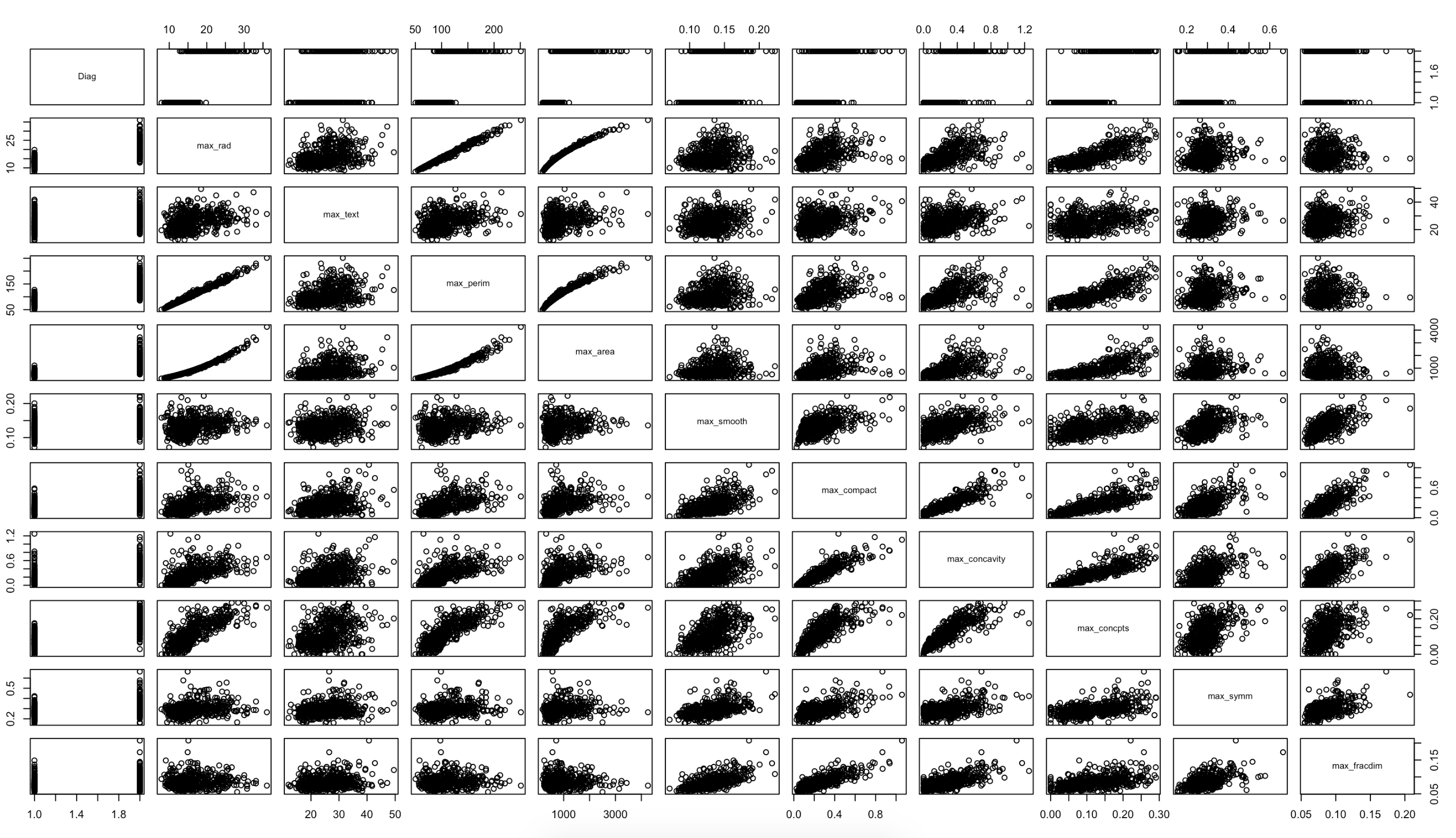
Correlation between “diagnosis” and the 2-11 variables:



Correlation between “diagnosis” and the 12-21 variables:



Correlation between “diagnosis” and the 22-31 variables:



# MODEL

First, we separate our data into training and test, 80% and 20%. For this type of dataset, we are going to create 2 models, a SVM Classifier and a Neural Network.

SVM CLASSIFIER

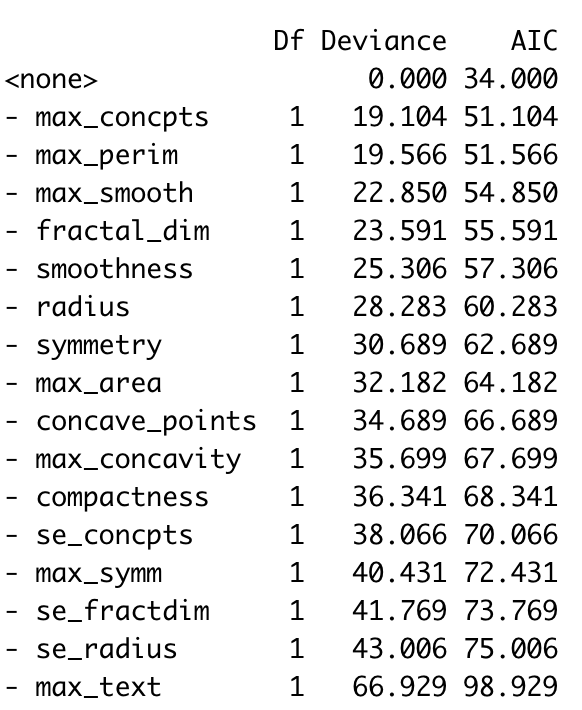
SVM is a supervised machine learning algorithm which can be used for classification or regression problems. It uses a technique called the kernel trick to transform the data and then based on these transformations it finds an optimal boundary between the possible outputs.

For our model we are going to use a “linear kernel” because is recommended to be used when we have large number of features.

The accuracy of the model for training data is 99.6 % and for the test data is 98.2 %.

Even we had good results we wanted to confirmed that were some options to improve the model:

The previous process selected 16 out of the 30 features:



With this 16 features we got an accuracy of 98.6% in the training set (1% less that our previous model) but we got the same accuracy of 98.2 % in our test set.

NEURAL NETWORK

# CONCLUSIONS