Capstone Project 2: Milestone Report I

Breast Cancer Prediction from Digitized Image of Fine Needle Aspirate (FNA) of a Breast Mass Using Deep Neural Network (TensorFlow 2.0)

1. Introduction

Breast cancer is a type of cancer with high mortality rates among women, and it is one of the most common causes of death in women. According to National Cancer Institute statistics in America, one out of eight women suffers from breast cancer and 6% of all deaths worldwide are caused by this type of cancer. Despite major advances in genetics and modern imaging, the diagnosis catches most breast cancer patients by surprise. For some, it comes too late. Later diagnosis means aggressive treatments, uncertain outcomes, and more medical expenses. Early diagnosis of breast cancer (maximum 5 years after the first cancer cell division) will increase survival chances from 56% to 86%. Besides, accurate diagnosis of breast cancer is of prime importance. Thus, a precise and reliable system is essential for the timely diagnosis of benign or malignant breast tumors.

In order to detect breast cancer, radiologists conduct Fine Needle Aspirate (FNA) procedure of breast tumor. FNA is a simple, inexpensive, noninvasive and accurate technique for detecting breast cancer. This procedure reveals features such as tumor radius, area, perimeter, concavity, texture and fractal dimensions. These features are further studied by medical experts to classify tumor as benign or malignant. Pathologists require a lot of expertise and skill to perform the analysis on the FNA sample. Applying the suitable features of the FNA results is the most important diagnostic problem in early stages of breast cancer. Hence, development of algorithms which rely on digitized image analysis, is of great interest. The key objective of this project was to predict breast cancer as benign or malignant using data set from the digitized image of FNA sample. We proposed to use data and build a Deep Neural Network (DNN) model using TensorFlow 2.0 that can help doctors find the cancer cells and ultimately save human lives. Doctors and pathologists in different Hospitals can use such a model to identify breast cancer from digitized images of FNA sample, which as a result would help them make early clinical decisions with greater accuracy.

2. Data Acquisition and Cleaning

The data set was acquired from the Kaggle Competition. This data (Breast Cancer Wisconsin Diagnostic Data Set) contains information computed from a digitized image of FNA of a breast mass along with labels (B=Benign, & M=Malignant). It describes the characteristics of the cell nuclei present in the image. We used this digitized image data in order to fit a better model. Since the data is already labeled, DNN using TensorFlow 2.0 is a perfect choice to build a predictive model with greater accuracy. The dataset has 569 observations and 33 columns. Out of the 33 total columns, ten were real-valued features that are computed for each cell nucleus:

The mean, standard error and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius. The column ‘diagnosis’ is the target variable that has two classes labeled as ‘B’ and ‘M’ which stands for Benign and Malignant respectively. The rest two columns; ‘id’ and ‘Unnamed: 32’ are dropped from the dataset as they are not relevant for our analysis. Our dataset didn’t have any missing value. Moreover, all the features are of data type 'float64' except the variable 'diagnosis' which is 'object'. In order to make it suitable for our Deep Learning algorithm, we replaced the ‘B’ and ‘M’ classes of the target variable by 0 and 1 respectively. As a result, the data type of our target variable became ‘int64’.

Our cleaned data has 569 observations and 31 columns. Our target variable is ‘diagnosis’, which is a binary variable with 0 representing Benign (not cancerous) and 1 representing Malignant (cancerous). The total number of samples that are benign are 357 (62.7%), while that of malignant/cancerous is 212 (37.3%). The proportion of malignant in our dataset is unusually large as compared to the real-world proportion of malignancy in which most masses are benign. However, this larger proportion of malignant cases is so important for our DNN as it learns well when data has a balanced proportion of classes.

3. Data Exploration

From the descriptive statistics, we can understand that the mean of 'smoothness_mean', 'compactness_mean', 'symmetry_mean', 'texture_mean', 'symmetry_mean', 'symmetry_worst', 'smoothness_worst', 'texture_worst' is almost same to the median represented by the 50%. However, the mean of the rest features is different from the median. Besides, there is remarkably large difference between the maximum value and the 75 percentile (75%) of the feature 'perimeter_mean', 'area_mean', 'area_se', 'perimeter_se', 'perimeter_worst', and 'area_worst'. In addition, there is also a big difference between 95th percentile and maximum value for 'perimeter_mean','area_mean', 'area_se', 'perimeter_se', 'perimeter_worst', and 'area_worst'. Hence, we can conclude that these features ('perimeter_mean','area_mean', 'area_se', 'perimeter_se', 'perimeter_worst', and 'area_worst') have outliers. There is no big difference between the 5th percentile and minimum value for all the variables.

Outliers are data values that are far away from other data values. It is important to check for the presence of such values in all the features as they are prone to affect our prediction. A box plot (or box-and-whisker plot) shows the distribution of quantitative data in a way that facilitates comparisons between variables. The box shows the quartiles of the dataset while the whiskers extend to show the rest of the distribution. You can see IPython notebook for further information on the box plot of the 30 features. The black dot above or below the whiskers indicate the presence of outlier. All the five features our data set have outliers. Specifically, while all features have only extremely large values (one direction), the 'smoothness_mean' feature has both extremely large and extremely small values as indicated by the dots in both directions. With that in mind, we also plotted a ‘*distplot*’ to check the skewness of the distribution of all the thirty features. Fig. 1 shows the distribution of all the thirty features.

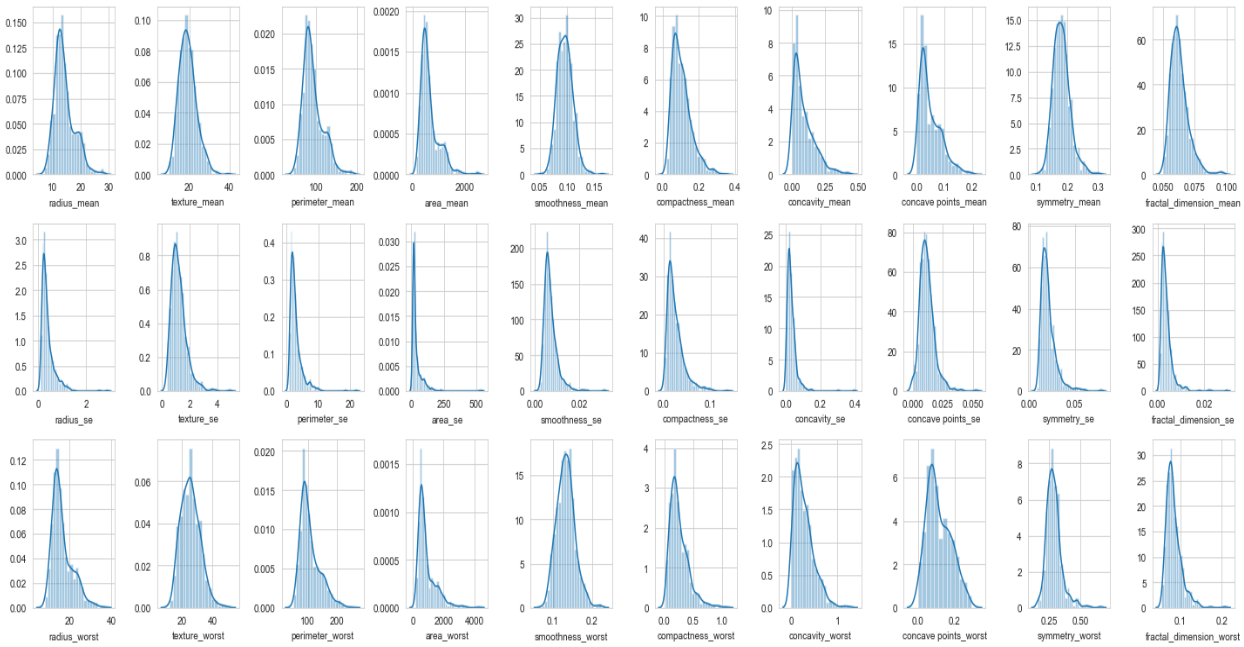

Figure 1: Distribution of all features.

The above figure depicts that 'texture_mean', 'smoothness_mean', 'symmetry_mean' 'texture_worst', 'smoothness_worst', 'symmetry_worst', 'concave points_se' appear to be normally distributed. It is also witnessed by the box plot displayed in IPython notebook that the median (middle line) of the box lies at the center between q1 and q3. On the other hand, the rest features are slightly skewed to the right.

Generally, 'radius_mean', 'perimeter_mean', 'concavity_mean', 'concave points_mean' are the four features that seem most relevant features for our prediction, followed by 'area_mean', 'compactness_mean', and 'texture_mean'. We observed Higher mean values in malignant group than benign group for all these six features. Contrarily, the fractal dimension mean, is the one that didn’t show any difference, while symmetry mean, and smoothness mean only show slight difference between the two groups. On the other hand, in most of the features the malignant groups have scattered data points as compared to the benign groups data points.

The ten features displayed below are the standard errors for the for the ten real valued features. Features like 'radius_se', 'perimeter_se', 'area_se', 'compactness_se', 'concavity_se', 'concave points_se' showed a slight difference in standard error between malignant and benign groups. The rest didn't show any difference in standard error. Fig 12 shows distplot of all the standard error related features.

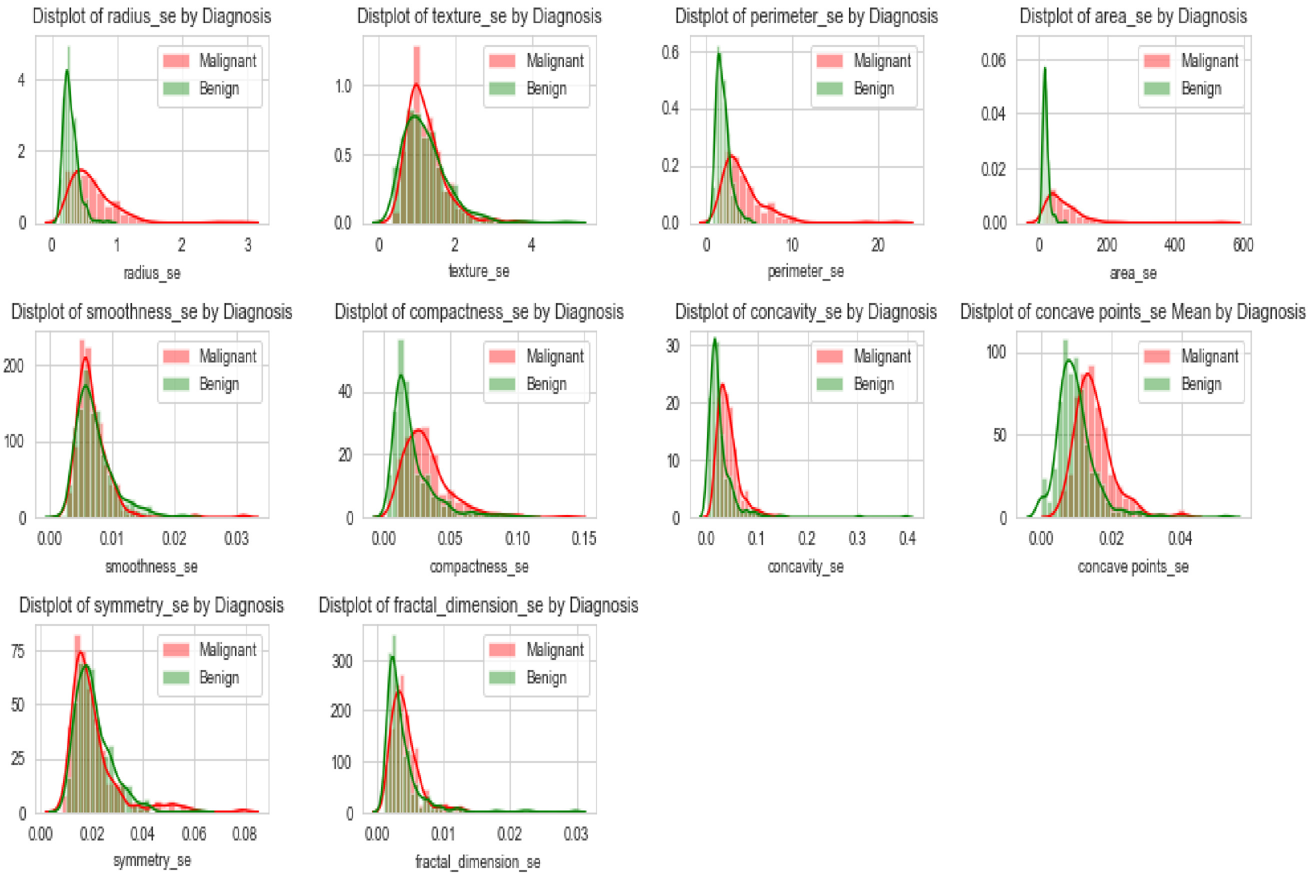

Figure 12. Standard error of all the ten real valued features by diagnosis

In our case, the suffix 'worst' indicates mean of the three largest values of each of the ten real valued features. The above chart indicated that the malignant group has a higher mean than the benign group for all the ten features. Features like 'radius_worst', 'perimeter_worst', 'area_worst', 'compactness_worst', 'concavity_worst', 'concave points_worst' showed bigger difference in mean between the malignant and benign group. Fig. 13 shows distplot of all the worst mean related features.

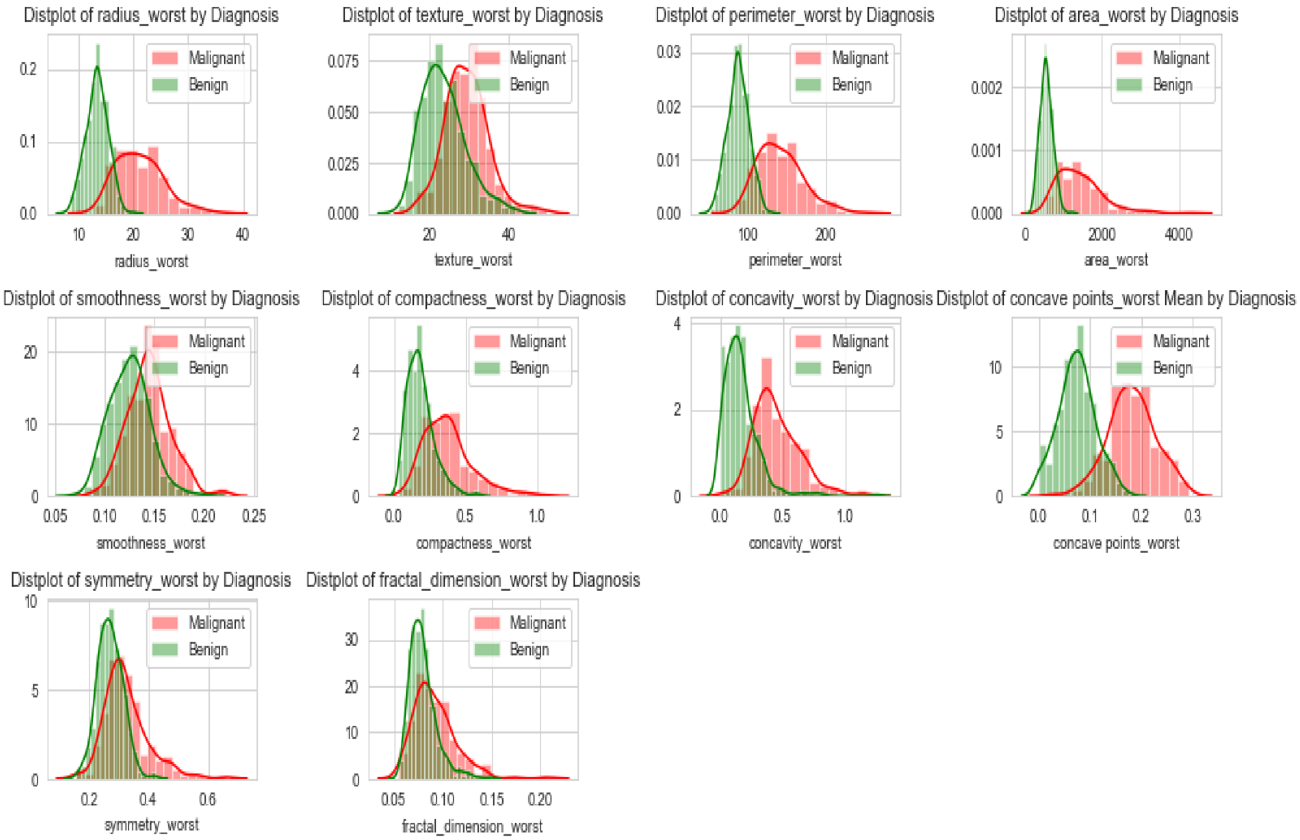

Figure 13. Worst mean of all the ten real valued features by diagnosis

The relevance of the six features stated above in classifying the target variable is witnessed by the scatter plot below. On the other hand, the scatter plot showed 'radius_mean', 'perimeter_mean', 'area_mean' are highly correlated for each other indicated by the steep data point among each other. For more information about correlation, we will do correlation matrix below. Fig. 14 shows pair plot of the six real valued features highly correlated with the target variable.

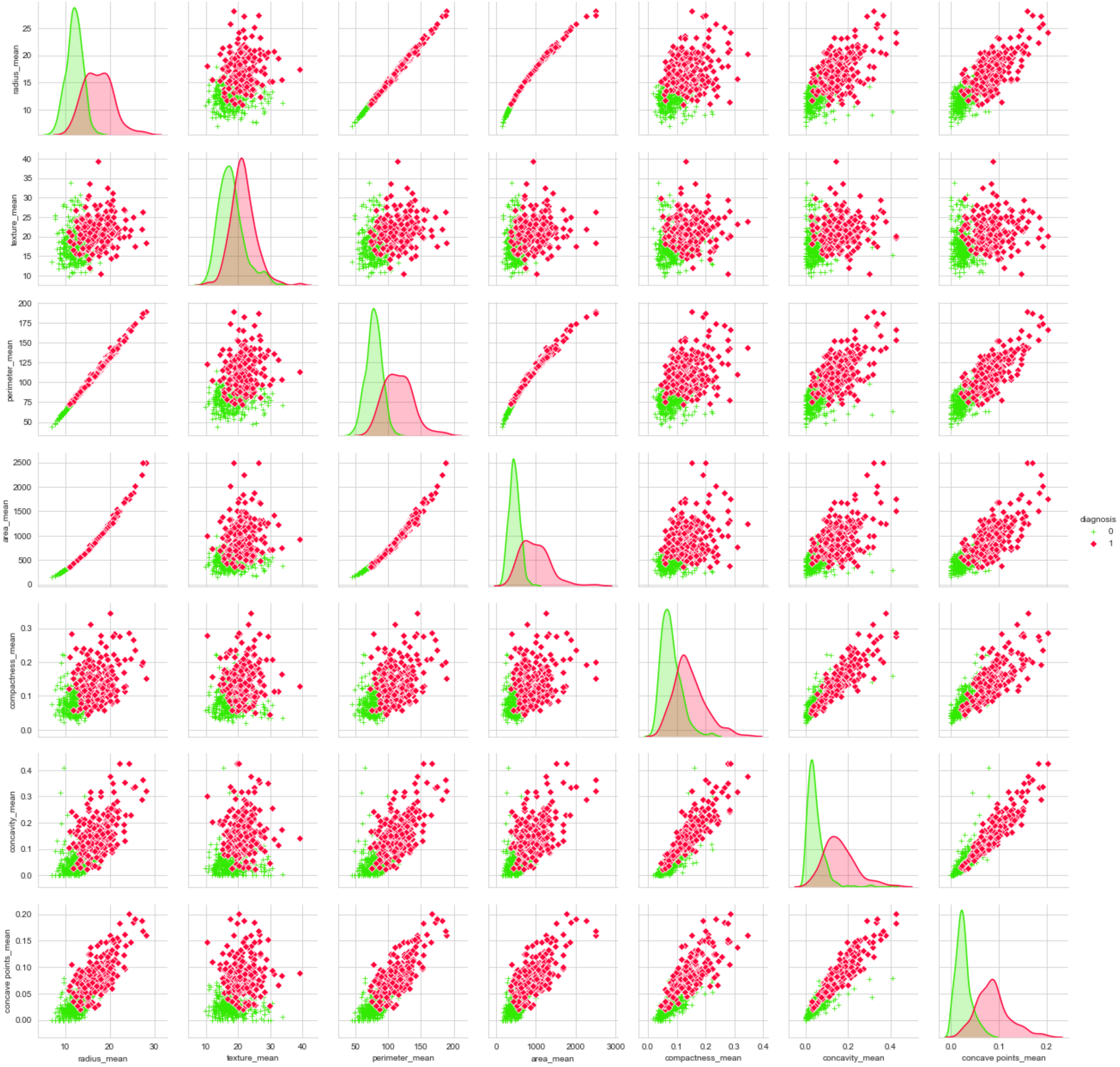

Figure 14. Pair plot of the six features by diagnosis

3.1 Correlation

The correlation matrix below shows the correlation of features among the top 20 highly correlated variables with the target variable.

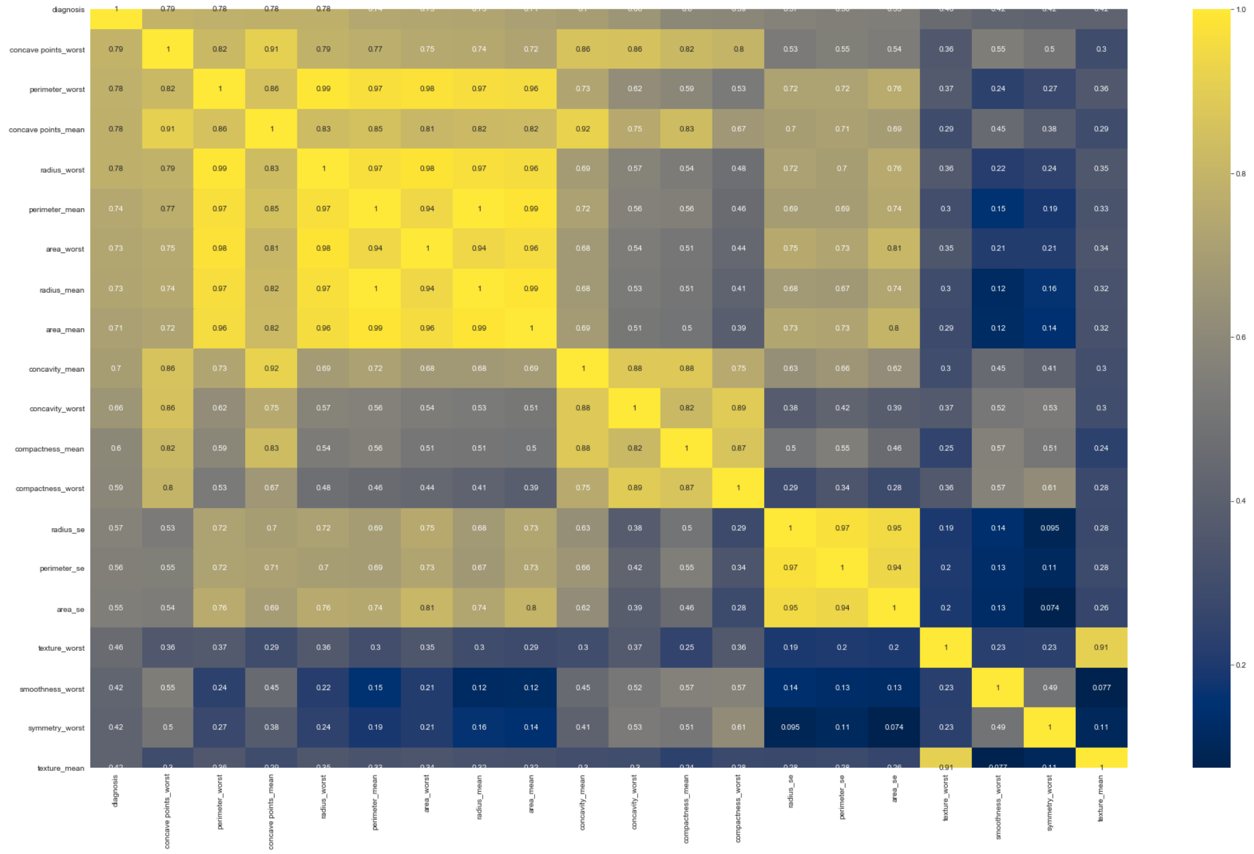

Figure 15. Correlation matrix of the top 20 correlated features to the variable diagnosis

Regarding the correlation with the target variable, concave points_worst, perimeter_worst, concave points_mean, radius_worst are the four top features that have the highest correlation with the target variable. On the other hand, the following features are the most correlated features with each other:

- radius_mean, perimeter_mean,area_mean,perimeter_worst, radius_worst, and area_worst
- texture_mean , and texture_worst
- area_se, perimeter_se, and radius_se

- compactness_mean, compactness_worst, concavity_mean, concavity_worst, and concave points_worst.

4. Outliers Handling

Data values that are either less than the 2nd percentile or above the 98th percentile are removed from the dataset. We removed four observations that lies in the mentioned range. Out of the four outliers that we removed, two were malignant and the other two were benign.