Classification of tumors into 2 categories (benign or malignant) based on features of tumor cell nuclei

Author: Rehan Ali

Capstone project 1 for SpringBoard Data Science Career Track

Mentor: Ryan Herr

Problem and the need for this analysis

A tumor can be malignant (cancerous) or benign (non-cancerous). This classification is made by a pathologist who examines the tumor cells obtained from a biopsy under a microscope. One method employed by pathologists is to examine the shape of a cell nucleus and make a judgement based on it.

The problem that is encountered in this regard, is that the shape is defined by many parameters such as the area of a nuclei, it's perimeter and numerous other numerical parameters. Making a judgement on changes in which of these parameters are more important relative to changes in other parameters requires a quantitative analysis. For example, one simple question to ask is if there is an increase in area of a cell nuclei but it is still circular in shape than what decision should a pathologist make?

The Client

The analysis described in this study is targeted towards cancer pathologists and oncologists. A pathologist's decision on tumor classification determines the type of treatment that an oncologist will undertake. The described analysis will aid a pathologist in deciding the type of tumor. Therefore, it will also increase the confidence of an oncologist who relies on the report of a pathologist.

Data used in this study

The data for this study comes from UCI machine learning repository and is created by Dr. William H. Wolberg of University of Wisconsin Hospitals. It is obtained from the following Kaggle dataset link:

https://www.kaggle.com/uciml/breast-cancer-wisconsin-data/data

The data is a comma separated file with rows labelled with patient identification number and the first column, labeled diagnosis, contains a series of ‘M’ (malignant) or ‘B’(benign) strings. The diagnosis (malignant and benign) will be referred to as classes. The rest of the columns contain values for 30 different parameters describing the shape of the cell nucleus. These are the features of the data.

Methodology and results

The methodology used and the results obtained are described here. It is suggested to read this along with the ‘Tumor classification.ipynb’ notebook as this closely follows this notebook file.

Firstly, the required libraries were imported. These included the following:

Libraries for importing and manipulating data:

- pandas
- numpy

Libraries for generating plots and performing statistical analysis:

- pyplot from matplotlib
- seaborn
- stats from scipy

Libraries for building and testing predictive models:

- decomposition from sklearn
- KNeighborsClassifier from sklearn.neighbors
- LogisticRegression from sklearn.linear_model
- GridSearchCV and train_test_split from sklearn.model_selection
- classification_report from sklearn.metrics

The data file named ‘Cancer.csv’ was read using pandas ‘read_csv’ method and the index column was set to the patient identification number. We will use this to make sure that no sample is repeated. Next, the head of the data was explored. The data frame showed 32 columns. One of these was 'diagnosis' which is the target variable. Another was 'Unamed: 32' column which contains only NaN values. The data also had columns with names ending with 'worst'. These are the worst estimates of different features and we do not need them.

Based on exploration described above the column named ‘Unamed: 32’ and the columns which had names ending with ‘worst’ were removed from the data frame. Also, it was made sure that no row had the same patient id number (the index column of the data frame). This ensured that no data was duplicated. The data frame also showed that different columns had different scales. Therefore, to visualize the data on the same axis the data was normalized. Also, the target variable was set as the index column.

Now we were ready to start performing analysis on the data. The first thing that was checked was correlation between different features. This was done because one can predict this even with out performing a correlation check as area and perimeter will be correlated with the radius. Moreover, if correlation between different features was found it would be important to know when we build predictive models for the data. We used a custom function 'plot_corr' for plotting correlation between different variables.

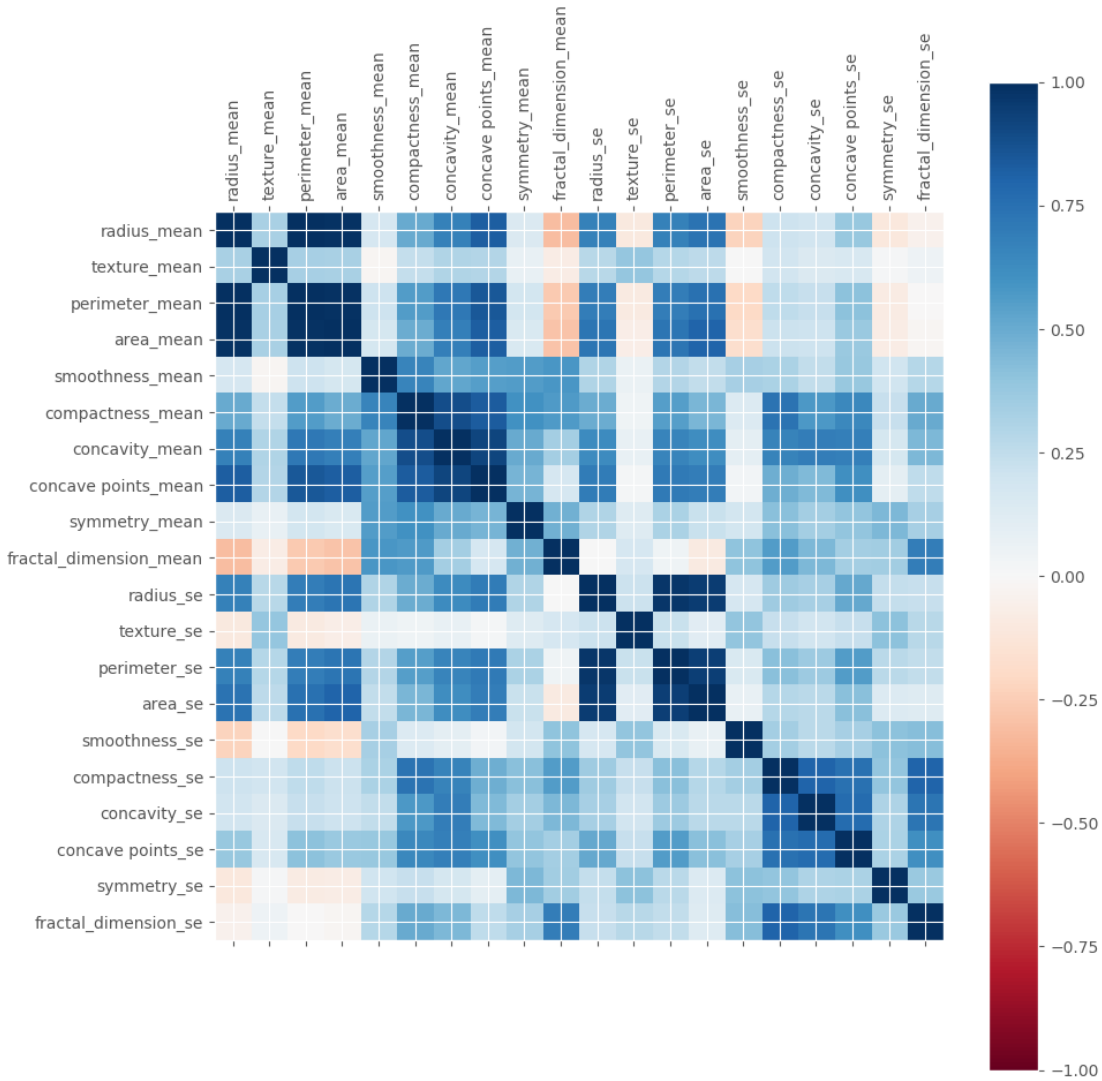


Figure 1: Grid showing correlation between different features. Blue means positively correlated and red means negatively correlated.

As was expected, there is a high correlation between radius, area and perimeter. Moreover, a high correlation between between compactness, concavity and number of concave points can also be seen. The take home message from this figure is that performing Principal Component Analysis (PCA) before building our predictive models might be beneficial.

Next, the differences between the features for the two classes were visualized using box plots overlaid with strip plots (individual data points). To achieve this in a simple way the data frame was transformed by melting it and then seaborn was used with 'hue = diagnosis'. The results are shown in the following figure:

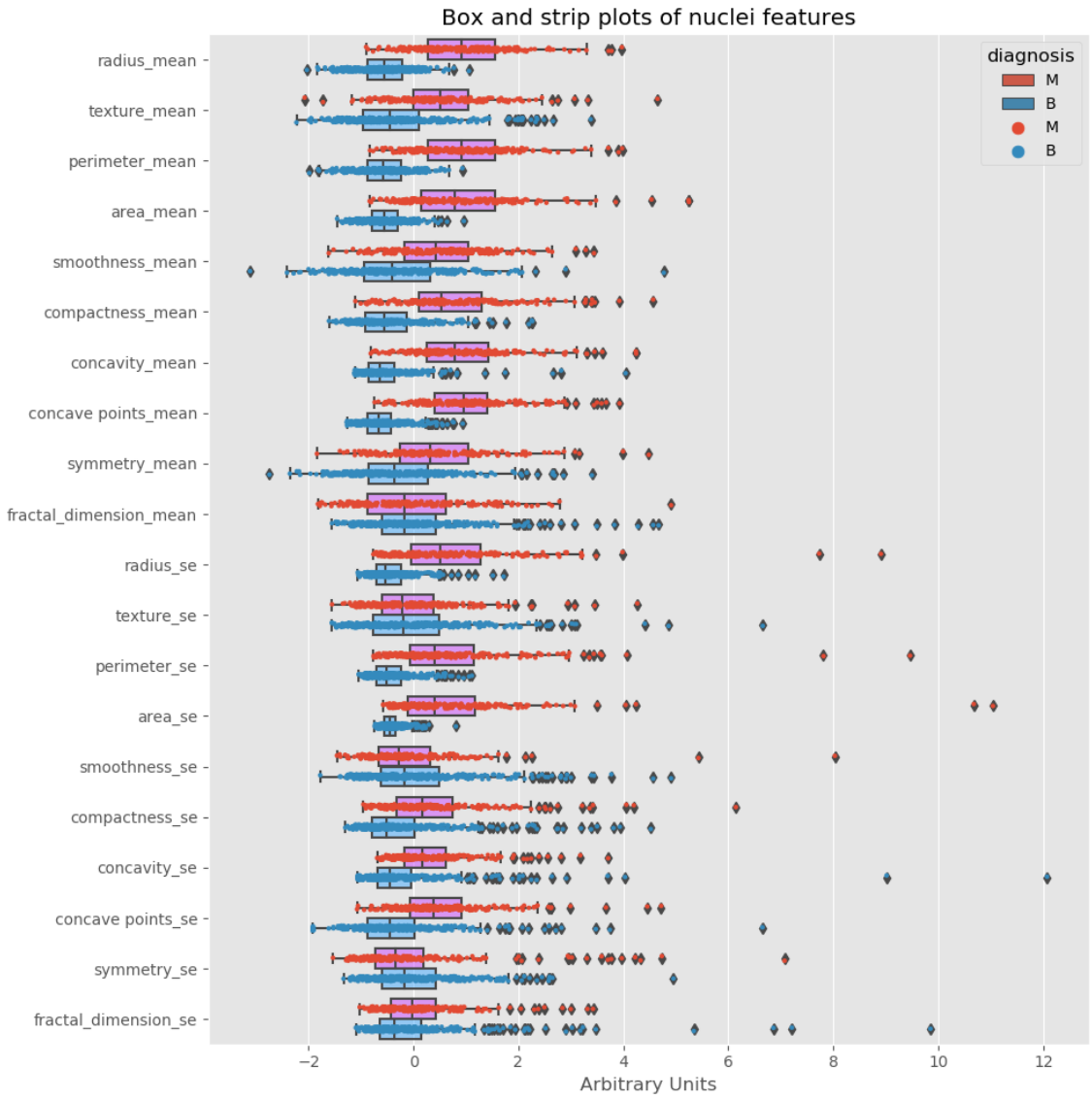


Figure 2: Box plots overlaid with strip plots showing normalized values (on the horizontal axis) for different features (on the vertical axis) for the two classes.

This figure shows that almost all the features differ between the two classes. But we do not have a quantification of these differences. To quantify the differences, so that we can see which variables show the greatest changes between the two classes we calculated percentage changes and made bar plots for these changes.

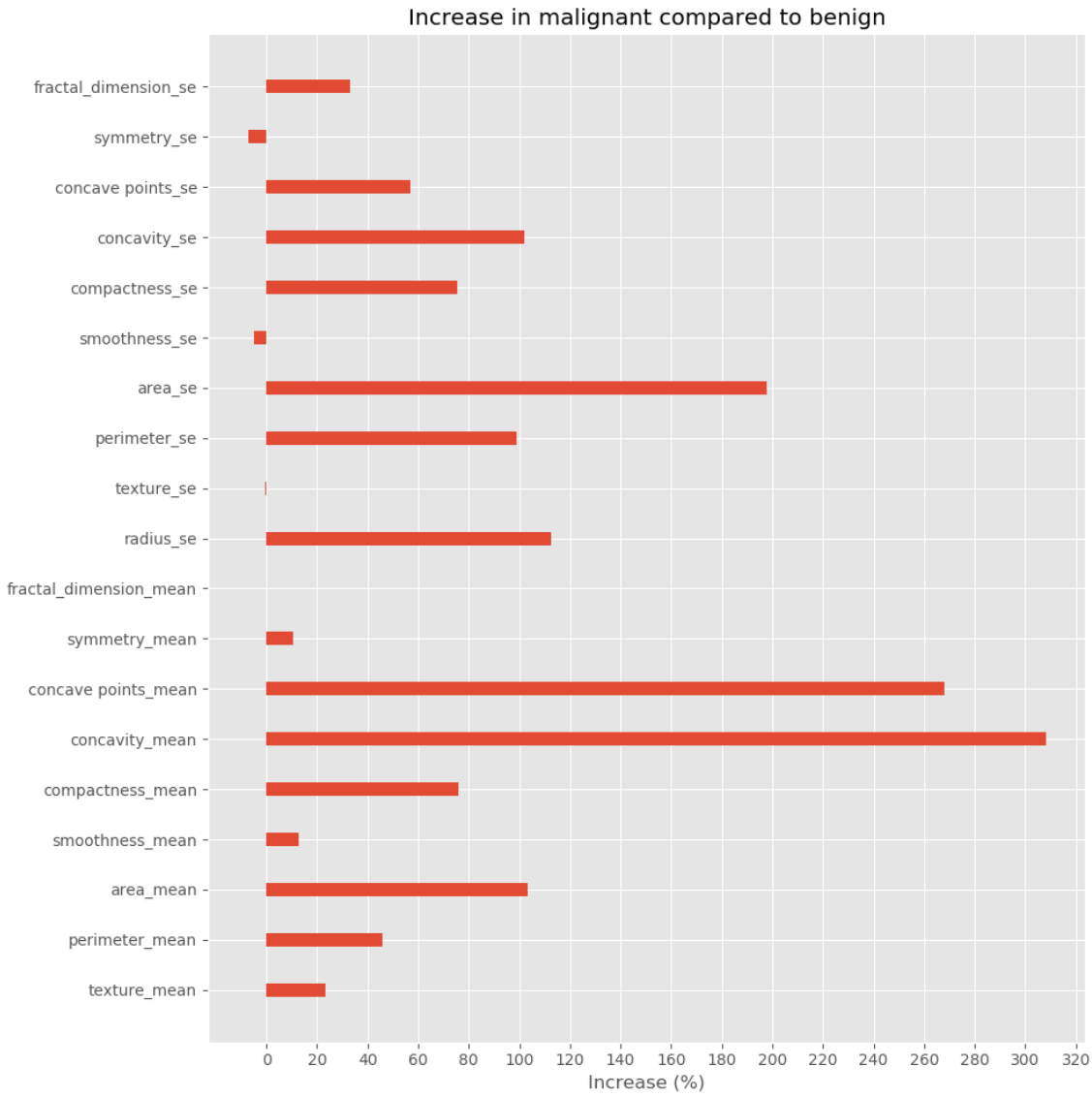


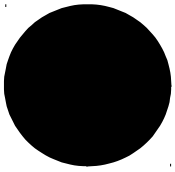
Figure 3: Percentage increase (on the horizontal axis) in the malignant class relative to the benign class for different features (on the vertical axis) is shown.

This figure provides us with more information about the changes in features between the two classes. We can see that the largest changes are observed in the concavity and the number of concave points. We can also see that the area of malignant nuclei also increases compared to benign nuclei. We already have enough information to build a coarse visual model for the two classes, which we will do shortly.

As the previous figures have shown, there is a difference in almost every feature of the nucleus for the two classes. To check if these differences were significant we performed a statistical test on the data. Firstly, it was checked if the data followed a normal distribution. The results showed that it did not follow a normal distribution. Therefore, mann-whitney u test was used to check for statistically significant changes. All the features are significantly different between the two classes (see the notebook for details).

Based on the results obtained, the following visual model is presented which describes the changes in the nucleus of a malignant tumor compared to a benign tumor.

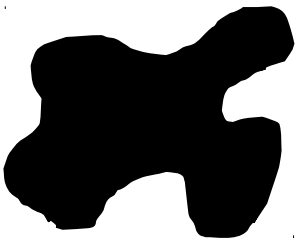
Benign nucleus



Smaller in size.

No distortion in shape

Malignant nucleus



Larger in size.

distortion in shape because of increase in the number of concave points and their concavity

Next we started building predictive models. Based on the fact that the features have correlations and the number of features is relatively large, the first logical step was to perform PCA to select features and reduce dimensionality of the data. The following figure shows the cumulative variance explained by different number of components.

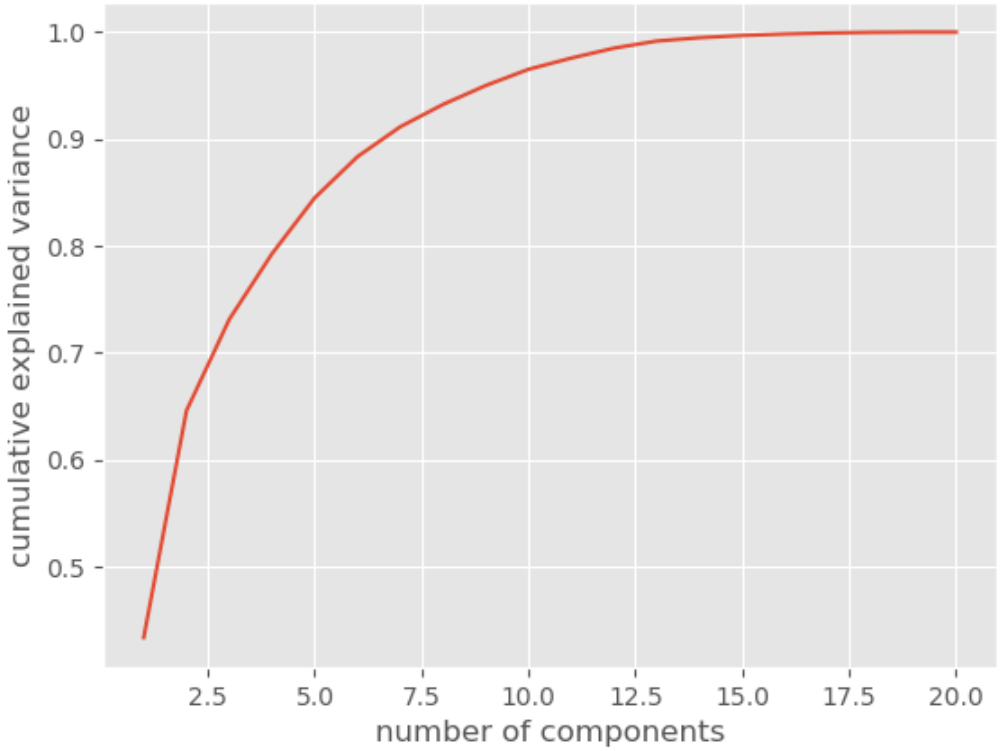


Figure 4: Cumulative variance explained by different number of components.

The figure above shows that variance of the data can be completely explained by 15 components instead of 20. Moreover, 10 components explain most of the variance in the data (~97%). We will use 10 as the number of components for building our models. We have reduced the dimensionality of our data by a factor of 2.

Using transformed data, split into test and train data set, the first model was built using k-NN method and grid-search-cv was used to tune the number of neighbors with cross validation. Then the model was fit to the training data and tested on the test data. The f-1 score for the model is 0.96.

The f1 score shows that the model predicts well on the test data but k-NN method has a few pitfalls: 1- If the test data is large, the model is slow to predict. 2- One cannot know the weight different features in the data have on predicting the class of tumors. Logistic regression, on the other hand, does not have these pitfalls. Therefore, a new model was built using logistic regression.

The f-1 score for the logistic regression model is 0.97. Not only will this model be faster in predicting on a large data set, it also has a better f-1 score compared to k-NN. But, the data used to train this model has been transformed using PCA. Therefore, it is not possible to know the effect of different features on predicting the tumor class. To know the effect of different features on predicting the tumor class, logistic regression was used to build a model on the training data without any data transformation. The notebook shows the coefficients for different features, which quantify the effect of different features on predicting the tumor class. But the f-1 score on test data is reduced to 0.95.

Based on the above discussion, the user can pick a model based on their own preferences. If a user needs the best predictive power then logistic regression with data transformed using PCA is recommended. But, if the user is more concerned about quantifying the effect of different features on predicting the tumor class logistic regression on non-transformed data should be used.

In summary, this report shows a systematic manner in which data should be analyzed. The data set used in this study has numerical features of two different classes of tumors but the workflow can be applied to any other data set which has numerical features. Also, different predictive models have been built and the advantages and disadvantages of each of these models explained. A pathologist can make an informed choice about the model that he or she needs to use. The next logical step in this study is to extract the numerical features of a nuclei shape directly from histology images using convolution neural networks and using the analysis described here to predict the class of a tumor. This goal is beyond the scope of this study, but once completed, the process of predicting a tumor class from just histology images can be completely automated.