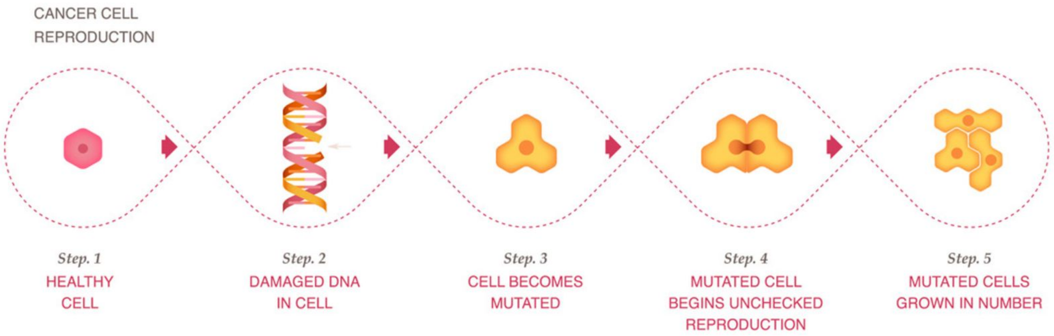
Machine Learning Nanodegree Capstone Project: Diagnosis of breast cancer

Version: v1.0
Authors: edgar.martinez.rico@gmail.com (Edgar Martinez Rico)
Status: Complete
Last Updated: 2017/06/01

I. Definition

Project Overview

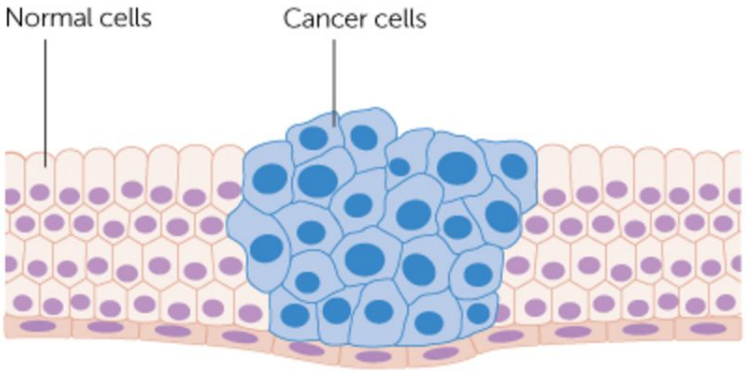
Cancer is a disease in which abnormal cells grow invading healthy cells in the body[1]. Cells constitutes the basic building blocks to build tissue and organs in the body. The process of cell growth could at times go wrong and lead into the development of new cells when the body does not need them. Or when old or damaged cells do not properly die. Therefore, the cell DNA is damaged and lost its instructions of how to divide[2]. This is referred as mutation and it means that a gene has been damage, as a result, the cell becomes mutated. The mutated cells starts rapidly reproduce and these cells grow in number. **Figure A** describes the cancer cell reproduction stages.



1

Figure A

This will result in the accumulation of cells that forms a mass of tissue that is referred as tumor, lump or growth. **Figure B** depicts a set of cancer cells.

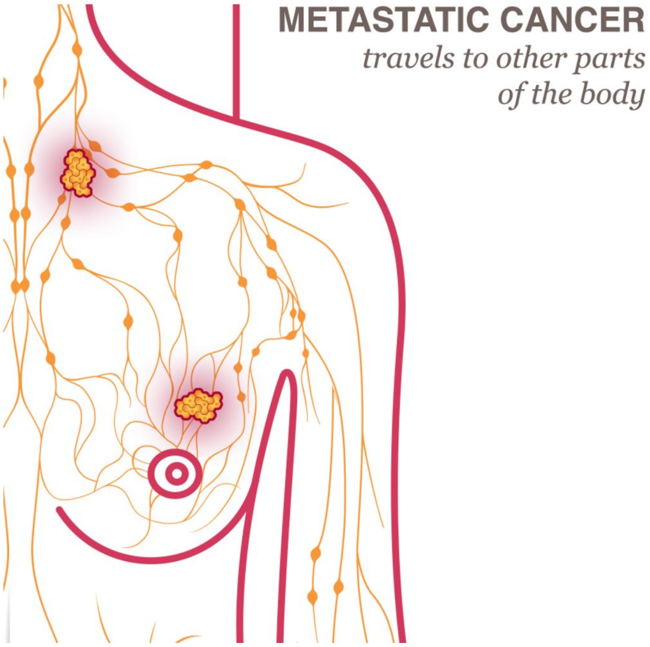


2

Figure B

Breast cancer begins with the formation of cancer cells in the breast that invade the nearby tissues or spread (metastasize **Figure C**) to other areas of the body. These cells can spread from the original malignant tumor if they entered into the blood vessels or lymph vessels.

[1] "What Is Cancer?" The National Breast Cancer Foundation.
[2] "What Is Cancer?" Cancer Research UK. 28 Apr. 2017. Web. 23 May 2017.



1

Figure C

The recognition of breast cancer as early as possible provides the better chance of treatment. As a result, knowing your breast is key, it is important to understand how your breast look and feel. Additionally regular mammograms and screening tests help to identify breast cancer in its early stage even before symptoms appear. The most common symptom is the development of a new lump or mass. This mass could be painless, tender, soft, rounded or have irregular edges which are more likely to be cancer.

Men breast cancer is rare, less than one percent of breast cancer are developed in men. According to the American Cancer Society breast cancer is the second leading cause of cancer death in women in the United States. The American Cancer Society estimates that 1 in 37 women will die from breast cancer[3]. On average, every 2 minutes a woman is diagnosed with breast cancer and one woman will die of breast cancer every 13 minutes. And according to the World Health Organization breast cancer is the most common cancer in women in developed and developing countries[4].

[3] “How Common is Breast Cancer?” American Cancer Society.

[4] "Breast Cancer: Prevention and Control." World Health Organization.

Problem Statement

The diagnosis of breast cancer is a critical step. In many cases this can be diagnosed with the use of a diagnostic mammogram, breast ultrasound or breast MRI. However, sometimes a biopsy might be required. There are two types of biopsies to diagnose breast cancer. Fine Needle biopsy removes cells from a suspicious tumor in the breast. Surgical biopsy provides the complete information in regards a lump and is the most accurate way to diagnose breast cancer.

However, the diagnosis with the fine needle biopsy has mixed success[5]. Different features such as training, experience and technical expertise of the physician performing the diagnosis[6]. Therefore, machine learning can be used for the diagnosis to determine if a lump is malignant or benign. This can be achieved with the use of a supervised learning technique that allow the classification of a lump in either benign or malign. The supervised learning model can be built taking in consideration different features of a cell nucleus such as radius, perimeter, area, compactness and symmetry to name a few.

Metrics

To evaluate the performance of the supervised learning model four measures will be used to assess the results:

		Predicted Breast Class	
		Malign	*Benign*
Actual Breast Class	*Malign*	TP	FN
	Benign	FP	TN

The measure above can be explained as follows:

- TP: A True Positive is when a malign tumor is predicted and is an actual malign tumor.
- FP: A False Positive is when a benign tumor is predicted as malign tumor.
- FN: A False Negative is when a malign tumor is predicted as benign tumor.
- TN: A True Negative is when a benign tumor is predicted and the diagnosis is benign tumor.

Based on the four measures described above and in order to evaluate the performance of the model using the following metrics will be calculated Accuracy, Precision, Recall and F1 score.

[5] Frable, William J. *Thin-needle Aspiration Biopsy*. Philadelphia: Saunders, 1983.

[6] "Cytopathology Palpation Guided Fine Needle Aspiration Analysis of FNA Procedure and Diagnosis" Pathology Outlines.

Accuracy

This perhaps is the most intuitive measure since it's the ratio of correctly predicted observations from the total number of observations. This is formally defined as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

However, this measure will be used for informative purposes only. Since the dataset is not symmetric, this measure is not a fit to evaluate properly the performance of the model.

Precision

Precision is the ratio of the correctly predicted observations from the total predicted positive observations. This is formally defined as follows:

$$Precision = \frac{TP}{TP + FP}$$

The higher the precision implies the decrease of false positive rate. For the model this measures the accurate predictions of malign tumors out of all the actual malign diagnoses.

Recall

Recall measure the ratio of the correctly predicted positive observations from all the observations in the actual class. This is formally defined as follows:

$$Recall = \frac{TP}{TP + FN}$$

For the model this measures the fraction of malign tumors that were predicted out of all the malign diagnoses.

F1 Score

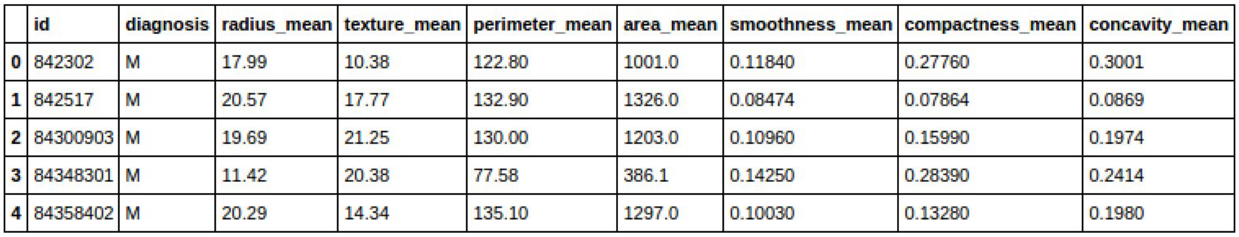
To evaluate properly the performance of the supervised learning model F1 score will be used. Particularly, if we take into consideration the uneven class distribution of the dataset. Since F1 score, measures the weighted average of precision and recall. As a result, this measure considers false positives and false negatives. This is formally defined as follows:

$$F1 = 2 \bullet \frac{precision \bullet recall}{precision + recall}$$

II. Analysis

Data Exploration

The data set for this project was obtained from Kaggle[7]. This dataset is divided into two main classes of tumors malign or benign. Each tumor is described using several features computed from digital images taken from a fine needle aspirate of a breast lump. These describes the characteristics of a cell nucleus in a three dimensional space[8]. The characteristics include radius, texture, perimeter, smoothness, compactness and concavity to name a few.



	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980

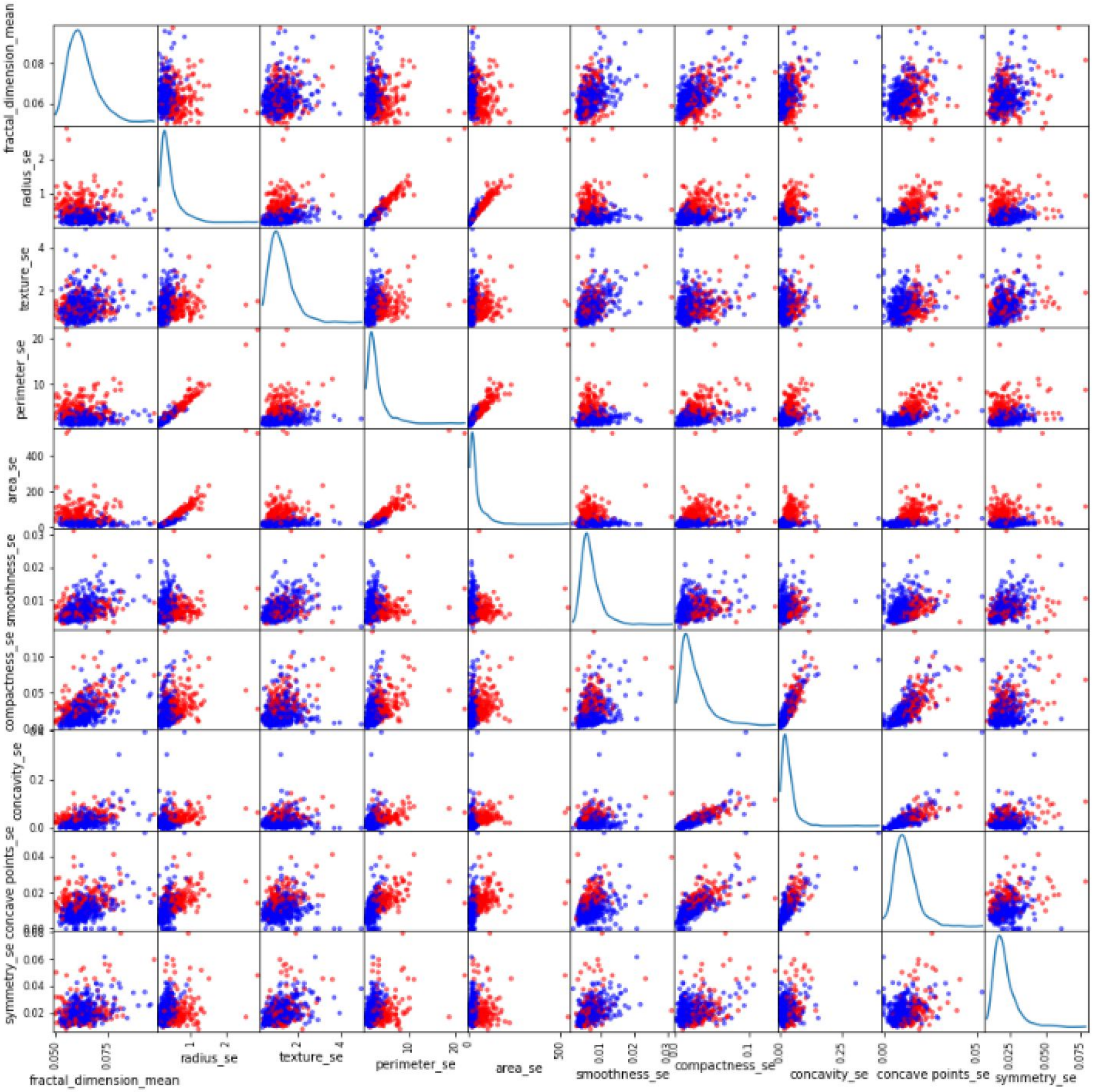
The data consist of 357 benign tumors and 212 malignant tumors. The tumor consist of 30 features in which all them are numeric except for the *"diagnosis"* feature. The feature called "*id*" does not provide value for the model since it represents a unique identifier. Similarly the feature "*Unnamed: 32*" does not have values. As a result, these features will be drop from the model. The target feature *"diagnosis"* has two values ***M*** for malign and ***B*** for benign. Such feature will be extracted and the remaining 27 features will be used for training and testing the model.

Exploration Visualization

To explore and digest the data, the use of scatter matrix plots will be presented. The following scatter matrix plot (**Figure D**) describes the distribution of the entire data points based on certain features (10 features). The data points in red represent malign lumps and in blue the benign lumps. The data points are being plotted in a scatter matrix to visually determine if there are linear correlation between multiple variables in the data. From the plot below we can identify that there might be correlation between "*radius_se*" and "*perimeter_se*" features. Another important aspect that can be observed from the plot is the data distribution. The diagonal of the scatter matrix plot describes the skewness of the data. From the plot is apparent that the entire features that are presented are left-skewed.

[7] *"Breast Cancer Wisconsin (Diagnostic) Data Set" Kaggle.*

[8] Bennett, Kristin, and O. L. Mangasarian. "Robust Linear Programming Discrimination of Two Linearly Inseparable Sets."



fractal_dimension_mean
0.08
0.06
radius_se
2
1
texture_se
4
2
perimeter_se
20
10
area_se
400
200
0
smoothness_se
0.03
0.02
0.01
compactness_se
0.10
0.05
0.00
concavity_se
0.4
0.2
0.0
concave points_se
0.04
0.02
0.00
symmetry_se
0.08
0.06
0.04
0.02
0.050
0.075
fractal_dimension_mean
1
2
radius_se
2
4
texture_se
10
20
perimeter_se
0
500
area_se
0.01
0.02
0.03
smoothness_se
0.0
0.1
compactness_se
0.00
0.25
concavity_se
0.00
0.05
concave points_se
0.025
0.050
0.075
symmetry_se

Figure D

Algorithms and Techniques

A supervised learning approach will be used to model the characteristics of a tumor. A classification model will be built to be able to determine if the characteristics provided of a tumor belongs to either benign or malign lump. This model should be able to generalize well enough for unseen data as a result, it should predict the corresponding output. Different supervised learning algorithms will be evaluated to determine the best model. Among the algorithms that will be evaluated are Gaussian Naïve Bayes, Support Vector Machine, Decision Trees and Random Forest Classifier to name a few[9].

Benchmark

Sklearn provides a Dummy Classifier which was used as a baseline. This classifier performs predictions using rules. This dummy classifier was used to compare the results with the real classifier model. The strategy that was used for the dummy classifier is *stratified.* This strategy generates predictions by respecting the training set's class distribution. The **F1 score** of this dummy classifier was obtained using the same set of training and testing sets as the real classifier. The best **F1 score** obtained with the dummy classifier was **53.06%**. **Figure E** denotes the confusion matrix of the best scores obtained by the Dummy Classifier.

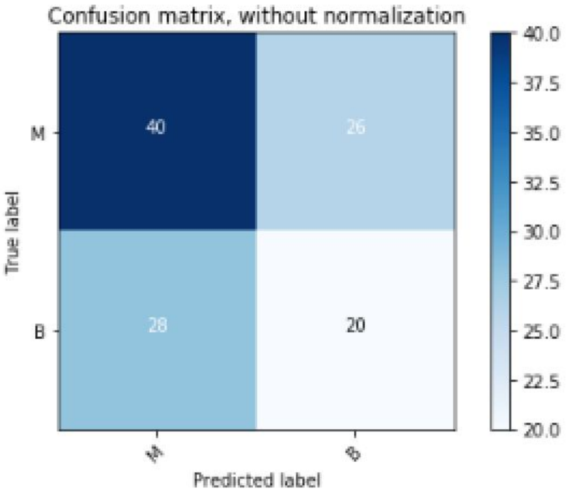


Figure E

Another approach is to compare prior results published on Kaggle. Most of the results published on Kaggle only considers accuracy as opposed to the **F1 score**. However, one of the project included a breakdown of the confusion matrix. Therefore, this project will be used as a second benchmark which obtained an F1 **score** of **98.33%**.

[9] *Choosing the Right Estimator — Scikit-learn 0.18.1 Documentation*.

III. Methodology

Data Preprocessing

Different steps were taken to prepare the data that was downloaded from Kaggle. These steps are described below:

1. Remove unnecessary features from the data. The data contains features that does not provide any value to the model. The features are: ***id*** and ***unnamed 32***.
2. Split the data into two different sets. This was performed by separating the tumor data into features and target column.
3. Outliers were not removed. Due to the nature of the data neither point was removed since each one represents a true diagnostic case.
4. Feature scaling of the data. Since the data is not normally distributed a non-linear scaling was applied. In this case the natural logarithm was applied to all the twenty seven features of the data. The **Figure F** shows the data after its preprocessing.

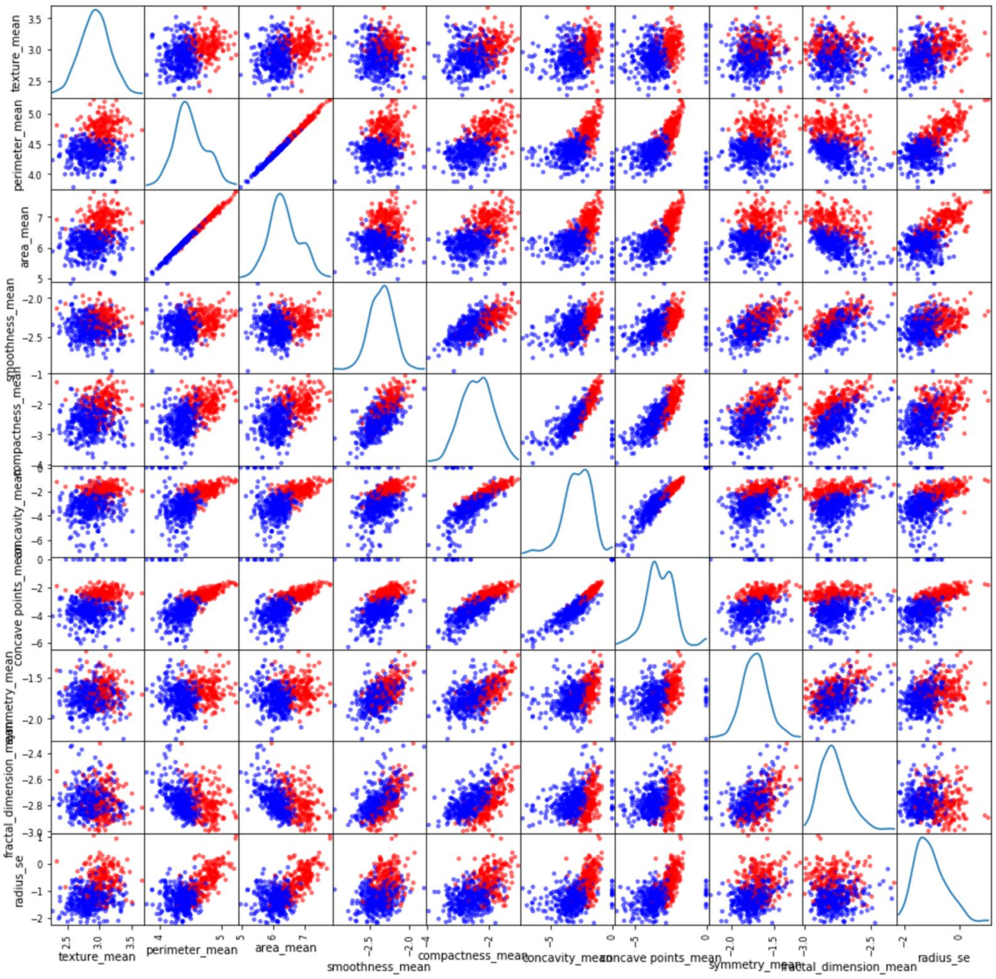


Figure F

Implementation

The implementation process can be defined into two main steps:

- Identify the principal components of the entire set of features.
- Apply several supervised learning algorithm and measure its performance.

PCA analysis

The tumor data set contains twenty seven features to describe each cell nuclei. To be able to determine the proper number of features for the model PCA was performed. PCA uses an orthogonal transformation that converts a set of possibly correlated features into a set of values of uncorrelated features called principal components[10]. The **Figure G** describes the PCA analysis performed for ten dimensions.

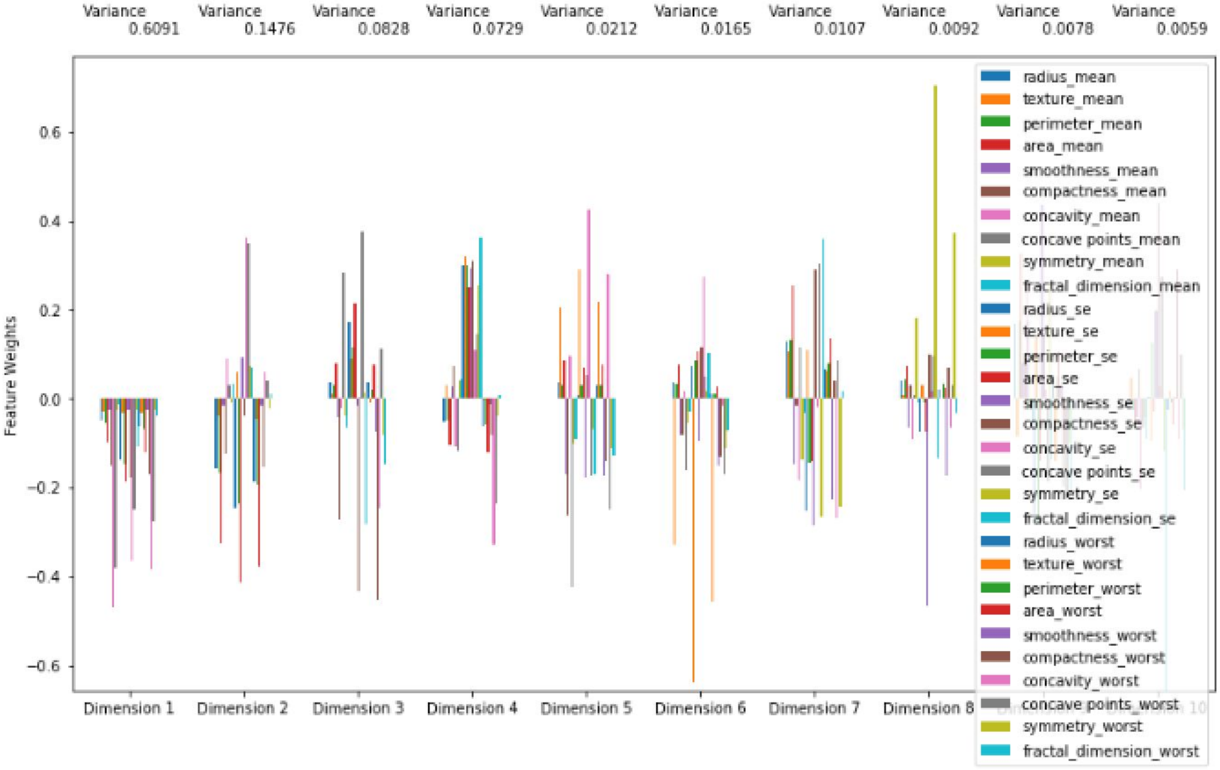


Figure G

From the principal components described on **Figure G**, it is apparent that the dimensionality reduction of the data will drastically reduce the number of features of the supervised model. The result of the principal component analysis can be interpreted as follows. The explained

[10] Principal Component Analysis. *Wikipedia*. Wikimedia Foundation, 24 May 2017. Web. 29 May 2017.

variance of the first dimension accounts for ~60%. Similarly the second dimension explains ~14%. However, it is clear that beyond the seventh dimension the explanation ratio falls below 1%. From this point becomes clear that additional dimensions does not justify the complexity of our model since the variance of Dimension 9 (0.0078) and Dimension 10 (0.0059) are too low. As a result, it was concluded that defining seven principal components will be defined in order to build the supervised model. These dimensions explain **~96.08%** of the entire data set.

Supervised Learning Algorithms

In order to select the proper supervised learning algorithm different algorithms were applied. The following supervised learning models that were used are available in scikit-learn:

- Support Vector Machine
- Gaussian Naïve Bayes
- Decision Tree Classifier
- K-Nearest Neighbors
- Random Forest Classifier

In order to train the classifiers the data used came after the PCA was applied. This data has seven features without including the target feature as described in **Figure H**.

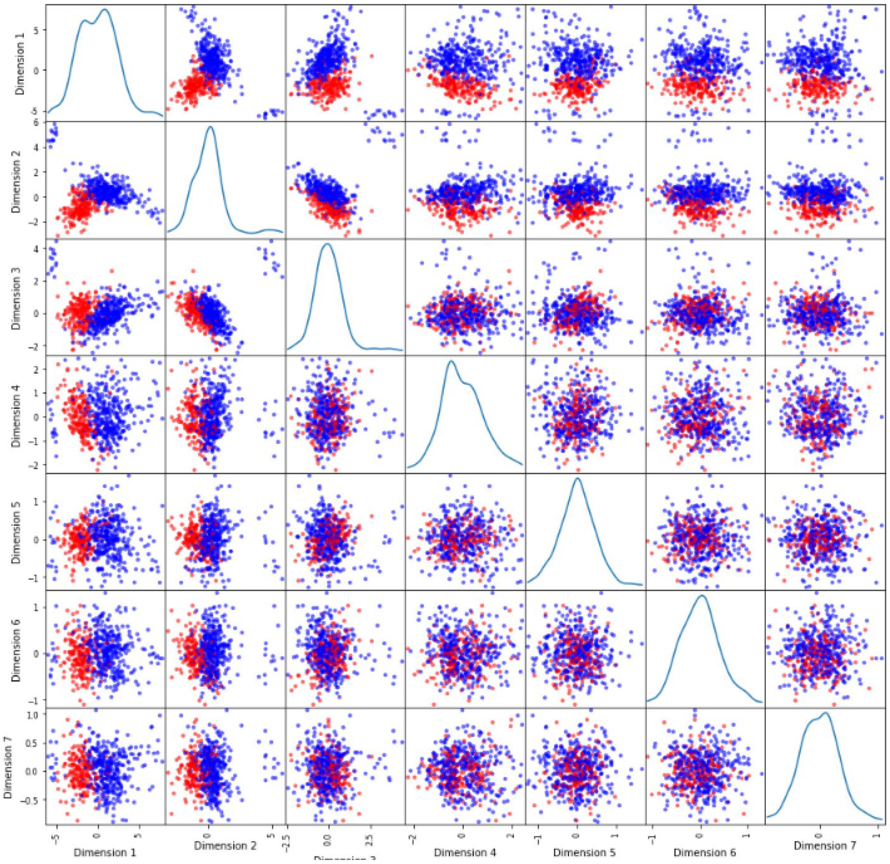


Figure H

To prevent over fitting of the supervised model part of the data was reserved to be able to test our model[11]. The data was split using the function train_test_split. The training data set consist of 455 training points (that represents approximately 80%) and the testing set contains 114 points (that represent approximately 20%).

The classifiers mentioned above were trained with the same training data set and also were evaluated against the same testing set. To be able to determine the right estimator each classifier was measure using **F1 score**. The following table describes the top three performing classifiers and its correspondent metrics.

`

	Support Vector Machine	**Random Forest Classifier**	**K-Nearest Neighbors**
F1 Score	96.77%	93.48%	95.65%

Although K-Nearest Neighbors produced a better outcome than Random Forest Classifier. The use of it was greatly discouraged mainly because in order to fine tune this algorithm there is no meaningful distance function (that can be used) due to the nature of the data and its sensitivity to localize data that might affect the outcome[12]. Therefore, support vector machines and random forest were further explored as the best candidates. This is described in detail in the refinement section.

Refinement

After the default model of support vector machine and random forest classifier obtained an F1 **score** of 96.77% and 93.48% respectively. Both models were fine tuned using an exhaustive search over different parameter values of each classifier. This was achieved using grid search and k-fold cross validation for each model. In terms, of k-fold cross-validation, this was specified to ten random splits to produce an average single estimation. This cross-validation was used in order to perform grid search for both classifiers. The grid search process of each classifier is discussed in detail on the subsequent sections.

[11] 3.1. Cross-validation: Evaluating Estimator Performance — Scikit-learn 0.18.1 Documentation. Web. 29 May 2017.

[12] K-nearest Neighbors | Brilliant Math & Science Wiki. Solve Problems in Math, Physics, and Algorithms. Web. 29 May 2017.

Support Vector Machine Fine Tune

For the support vector machine model few parameters were considered in order to improve the **F1 score** of the model. The following table describe the parameters and the values that were applied to perform the grid search.

Parameter	Meaning	Values
C	Defines the tradeoff of the decision boundary	1, 10, 100, 1000
kernel	Specifies the type of kernel to be used	linear, rbf
gamma	Defines the influence of a training sample.	0.01, 0.001, 0.0001, 0.00001

After applying grid to the support vector classifier the initial **F1 score** was slightly improved from 96.77% to **98.97%**. The fine tuned parameters of the final classifier are as follows:

Parameter	Value
C	100
kernel	rbf
gamma	0.01

Random Forest Classifier

For this classifier some parameters were considered improving the **F1 score** of the model. The following table describe the parameters and the values that were applied to perform the grid search.

Parameter	**Meaning**	**Values**
max_depth	The maximum depth of the tree.	3, None
max_features	The number of features to consider when looking for the best split	1, 2, 3, 4, 5, 6, 7
min_samples_split	The minimum number of samples required to split an internal node	2, 3, 7, 11
min_samples_leaf	The minimum number of samples required to be at a leaf node	2, 3, 7, 11
bootstrap	Whether bootstrap samples are used when building trees	True, False
criterion	The function that measures the quality of the split	gini, entropy

After applying grid to the random forest classifier the initial **F1 score** was greatly improved from 93.48% to **98.95%**. The fine tuned parameters of the final classifier are as follows:

Parameter	Value
max_depth	None
max_features	3
min_samples_split	2
min_samples_leaf	3
bootstrap	True
criterion	gini

The difference between the two fine tuned classifiers is minimal. The support vector machine **F1 score** is slightly above the random forest classifier by ~0.02%. Therefore, to define the best classifier the model is discussed in the following section.

IV. Results

Model Evaluation and Validation

Since the **F1 score** among the two fine tune classifiers are pretty close. Further analysis was required to determine the best model. This was determined by analyzing more the different components of the **F1 score**. The analysis was performed taking into consideration the confusion matrix of each model described on **Figure I** and **Figure J**.

Support Vector Machine

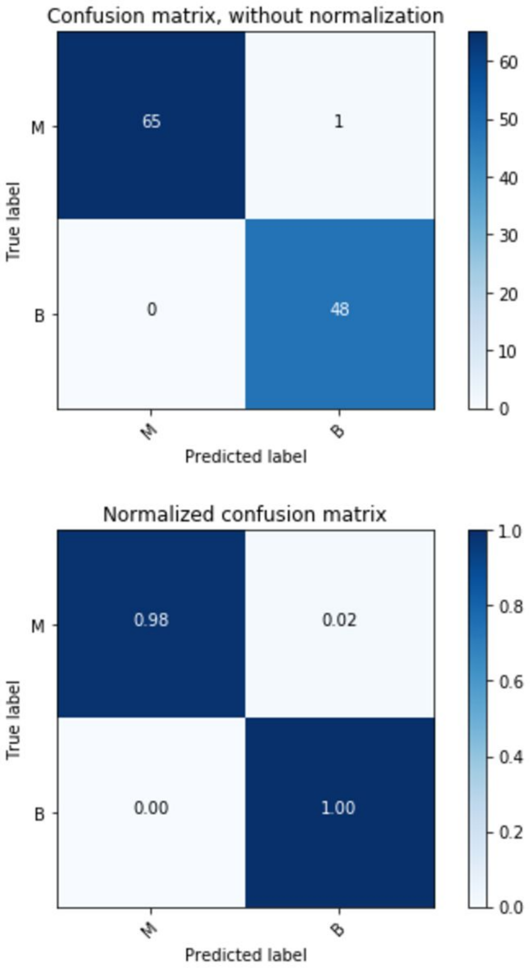


Figure I

Random Forest Classifier

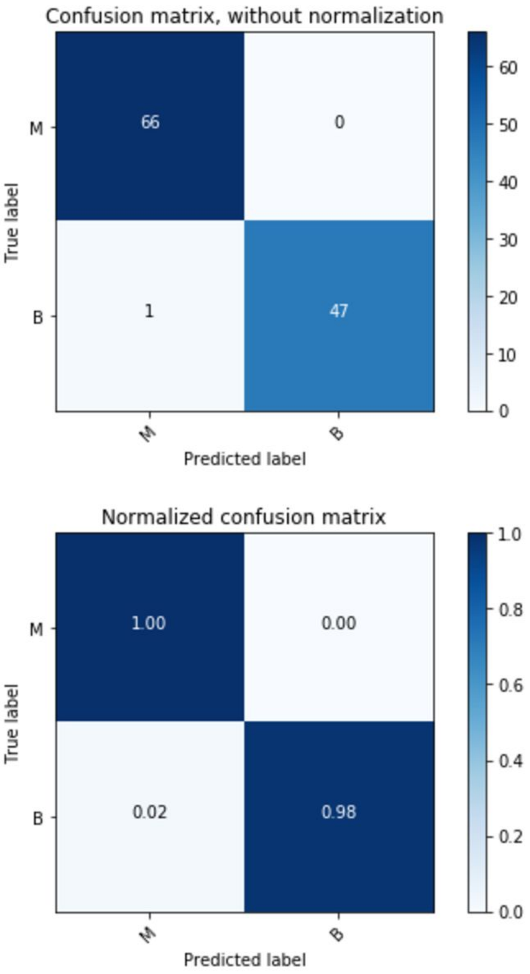


Figure J

The following table provides the different components of the **F1 score** that were analyzed to determine the best model.

Measures	**Support Vector Machine** **Values**	**Random Forest Classifier** **Values**
Accuracy	99.12%	99.12%
Precision	**100%**	**98.50%**
Recall	**98.48%**	**100%**
F1 score	98.97	98.95%

Because the goal of the project is to diagnose breast cancer (malign tumors) and since **F1 score** considers different measure. It is important to denote the model potential errors. These errors are known as type I and type II errors[13]. Type I error is associated with false positives; which relates to a benign tumor that is predicted as malign tumor. This ratio is capture by precision measure. A type II error is associated with false negative. Which occur when a malign tumor is predicted as benign tumor. This error type is measured by recall.

With respect to the diagnosis of breast cancer generally both errors types might lead to significant concerns. However, in terms of error type I, a second test might produce a conclusive result and be able to determine the lump in question is not malign. Unfortunately, this is not the same for error type II in which a treatment of breast cancer will not be recommend and will lead to fatal consequences (because the model concluded that the lump was benign when is not the case).

Due to the fact that error type II has larger implications it was apparent that the most important measure to consider was recall. Therefore, random forest classifier is the recommended model for breast cancer diagnosis because it provides the best recall rate.

[13] Type I and Type II Errors. *Wikipedia*. Wikimedia Foundation, 21 May 2017. Web. 31 May 2017.

Justification

Ultimately, the use of random forest classifier provides results that are consistently superior to the benchmark of the dummy classifier. The figure of the best **F1 score** obtained by the dummy classifier was **53.06%** that contrast with the **98.95%** that was obtained with the random forest classifier. Similarly, this score is slightly above from the score published on a project in Kaggle. Such project obtained an F1 **score** of **98.33%**. The overall improvement of the random forest classifier presented on this paper represents an increase by ~0.62%. To sum up, the results presented in this paper require further statistical test for significance, such as t-test. Unfortunately, this cannot be performed due to the small sample size of the results. However, the model presented in this paper represents a promise of a potential sophisticated diagnosis of breast cancer in the near future and how machine learning can be applicable in this domain.

V. Conclusion

Free-form Visualization

During the development of the project it became apparent that perhaps the best benchmark measure for cancer diagnosis is recall. Since in this domain it represents the ratio of false negatives. This have huge consequences if the ratio is too high for the model. Since, this implies that a malign tumor was predicted as benign. Hence, treatment of the tumor will not be recommended. In contrast, in the event that the precision rate is slightly behind, this could be addressed by performing a second diagnosis. This aspect should be discussed broadly with the right forum of matter of experts to determine the proper trade off between the precision and recall ratio.

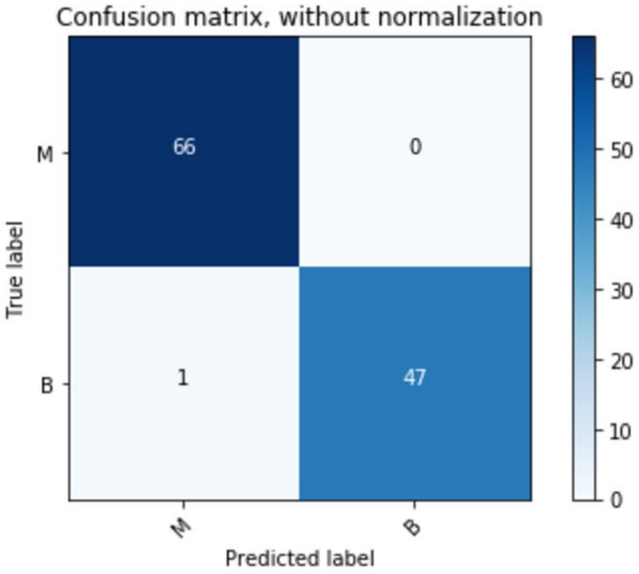


Figure K

Figure K denotes the confusion matrix obtained using the random forest classifier.

Reflection

I have always been fascinated about how technology can help to solve difficult challenges found in clinical science. And I strongly believe that machine learning is a powerful tool to help in the diagnosis of a terrible disease such as breast cancer. This project gave me the opportunity to explore this area.

One of the most difficult aspects of the project was to try to understand the different features of the dataset. The understanding of the dataset was a crucial step to define different techniques and algorithms. Therefore, the use of different plots were used to analyze further the dataset. The use of scatter plot matrix revealed that the data was not normal distributed. In order to compensate it feature scaling was applied to the dataset to reduce skewness in the data. Similarly, the use of principal component analysis was performed to reduce the number of features from 27 to 7. Using this data with different classifiers such as, support vector machine, decision trees, k-nearest neighbors and random forest (to name a few) were used to evaluate its performance. To determine the best performing algorithm recall measure was critical to determine the best model. The use of recall clearly pointed out that random forest classifier was the best model for the diagnosis of breast cancer.

The results provided in this paper clearly demonstrate the potential of machine learning in the diagnosis of breast cancer. However, further evaluation is required to be able to determine the correct model that can be use in real life.

Improvement

One of the obvious improvement that needs to be solved is the lack of data. The dataset that was downloaded from Kaggle is too small to drive relevant conclusions or further statistical analysis. In addition to it another model that would require further analysis is Neural Networks. However, this is not the only aspect that requires further refinement. One key aspect that needs to be addressed it the availability of the system presented here. To take full advantage of machine learning in this domain the model presented here should be able to scale. It means that it should be able to learn new data points over time and should be able to provide results that are readily available from a tablet, smartphone via web.

References

- Breast Cancer Wisconsin (Diagnostic) Data Set | Kaggle. Web. 15 May 2017.
- Bennett, Kristin, and O. L. Mangasarian. Robust Linear Programming Discrimination of Two Linearly Inseparable Sets. Optimization Methods and Software 1.1 (1992): 23-34. Print.
- Breast Cancer: Prevention and Control. World Health Organization. World Health Organization. Web. 14 May 2017.
- Choosing the Right Estimator. Choosing the Right Estimator — Scikit-learn 0.18.1 Documentation. Web. 21 May 2017.
- Frable, William J. Thin-needle Aspiration Biopsy. Philadelphia: Saunders, 1983. Print.
- General Cytopathology - Analysis of FNA Procedure and Diagnosis. Web. 15 May 2017.
- How Common Is Breast Cancer? American Cancer Society. Web. 14 May 2017.
- Nbcf. What Is Cancer? The National Breast Cancer Foundation. www.nationalbreastcancer.org. Web. 14 May 2017.
- 3.1. Cross-validation: Evaluating Estimator Performance: Evaluating Estimator Performance — Scikit-learn 0.18.1 Documentation. Web. 29 May 2017.
- K-nearest Neighbors Brilliant Math & Science Wiki. Solve Problems in Math, Physics, and Algorithms. Web. 29 May 2017.
- Principal Component Analysis. Wikipedia. Wikimedia Foundation, 24 May 2017. Web. 29 May 2017.
- Type I and Type II Errors. Wikipedia. Wikimedia Foundation, 21 May 2017. Web. 31 May 2017.