

# Data Mining: A Complete Analysis For Transformation, Validation and Evaluation Of An Applied Data Set

Mariana Ávalos Arce  
Universidad Panamericana  
Data Mining, Fall 2021



Guadalajara, Mexico

# Data Mining: A Complete Analysis For Transformation, Validation and Evaluation Of An Applied Data Set

Mariana Ávalos Arce  
Universidad Panamericana  
Data Mining, Fall 2021

## Abstract

The purpose of this document is to perform a thorough analysis of a data set used for a Binary Classification Problem, in order to make decisions on how to transform and validate/resample the data set before feeding it to a Machine Learning Problem (Logistic Regression Model), and after this is done, evaluate the resulting model with different metrics of performance: accuracy percentage, Cohen's Kappa coefficient and ROC AUC technique. The analysis includes some of the basics of data distribution and dispersion analysis of the said attributes. The document includes code snippets of some of the actions performed in order to achieve a better understanding of the results from the reader.

---

## 1 The Data Set

The data set used in the following pages is owned by UCI Machine Learning [7], and was found through Kaggle's data set browser. In simple words, this data set is used for a Supervised Classification Problem in which a Machine Learning algorithm learns to predict whether a female patient's breast tumor is malign or benign, given 10 attributes from said tumor. The features or attributes in the data set were extracted from digitized images of a breast cancer cell in different tumors, which describe the cell nuclei presented in the images. A brief description of each column or attribute is provided for clarity on further sections:

1. **id**: unique number of the patient.
2. **diagnosis**: M = malign or B = benign.
3. **radius\_mean**: mean of distances from center to points on the perimeter.
4. **texture\_mean**: standard deviation of gray-scale values found on the tumor's image.
5. **perimeter\_mean**: mean size of the core tumor.
6. **area\_mean**: mean area of the tumor.
7. **smoothness\_mean**: mean of local variation in radius lengths that represent smoothness.
8. **compactness\_mean**: mean of perimeter raised to the power of 2 over the area.
9. **concavity\_mean**: mean of severity of concave portions of the tumor's contour.
10. **concave points**: mean for number of concave portions of the tumor's contour.
11. **symmetry\_mean**: tumor's symmetry measure.
12. **fractal\_dimension**: mean for *coastline approximation* - 1.

### 1.1 Understanding The Data set

The first approach to a data set should always involve a glimpse of the kind of instances that are registered in the table. We will visualize the first 5 instances that appear in said data set using Pandas' `head()` function, which outputs the first 5 instances of a data frame when no argument is sent [11].

id	diag	radius	texture	perimeter	area	smooth	compact	concavity	concave pts.	symmetry	fractal
842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.30010	0.14710	0.2419	0.07871
842517	M	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.08690	0.07017	0.1812	0.05667
843009	M	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.19740	0.12790	0.2069	0.05999
843483	B	11.42	20.38	77.58	386.1	0.14250	0.28390	0.24140	0.10520	0.2597	0.09744
843584	M	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.19800	0.10430	0.1809	0.05883

Table 1: First five instances in the data set.

After seeing what appears on Table 1, we can conclude that the data set has the target column of the binary classification under the column `diag`, which stands for *diagnosis*, and it contains inside what appears to be either B or M, when the tumor in the row is either Benign or Malign, but we will confirm this with another function later in this section. Nevertheless, we can confidently say that all the attributes related to the target are floating point numbers, some of which have a large range of values, as in columns `area` and `perimeter`, especially in comparison to columns with a smaller range of values, which are pretty much all remaining columns: **the data set contains significant differences among the attributes' range of values**, which is one of the indicators for performing a Standardization Transformation. The column `id` is indeed numeric but let's not get fooled by the data type: this column brings no information regarding the prediction of the target column.

We can see that the `diagnosis` column is the target column that contains the classification for each tumor row, but how to know for sure that this is a data set destined to predict for a Binary Classification Problem? Well, if we consider that a Binary Classification Problem consists in a set of attributes that can either belong to one class or another, it is necessary to have **only two** possible values in the target column. For confirming this, we will use the function `groupby()`, which involves some combination of splitting the object, applying a function, and combining the results [12]:

```
df.groupby('diagnosis').size()
```

This snippet is definitely not the only code necessary, but the current document assumes the reader knows the missing previous steps of including libraries and reading a .csv file as a dataframe. Back to the point, we get the results presented in Table 2.

diagnosis	
Value	Count
B	357
M	212
dtype: Object	

Table 2: Results from `groupby()` function applied to *diagnosis* column.

As we were expecting, the Table 2 shows that the target column `diagnosis` contains only two possible values: B or M, which stand for Benign and Malign. Thus, each of the rows is a tumor and the column `diagnosis` describes the final classification of said tumor. Additionally, the output includes the `dtype` of the column, which is the data type(s) that the column contains [11]. As described in Pandas Documentation [11], the Pandas library is based on Python's numeric library Numpy, and therefore, Pandas identifies any data type that is not numeric as `Object`, when in reality, the column is of `string` type. For further analysis, this column will be binarized in following sections, so that further Machine Learning models can use the column. Most Machine Learning algorithms for Binary Classification need that **all attributes are numeric**, which we can check by consulting the `dtypes` attribute of our data frame.

```
df.dtypes
```

The attribute `dtypes` of a data frame shows all the data types by column of said data frame [11]. The snippet above outputs the results presented in Table 3.

Column	Data Type
id	int64
diagnosis	object
radius	float64
texture	float64
perimeter	float64
area	float64
smoothness	float64
compactness	float64
concavity	float64
concave points	float64
symmetry	float64
fractal dimension	float64

Table 3: Results from `dtypes` attribute of the present data frame.

Considering Table 3, we can say that the data set contains all attributes in a numeric data type, more specifically, in a floating point number. Now, by checking the column name and its data type, each row is a collection of different measures taken from the tumor of each row and registered its measures in different units, which explains the difference in each column's ranges in value.

Coming back to the dimensions presented in Table 2, we can get an idea of the overall dimensions of the data set by consulting the data frame's `shape` attribute, which tells us the matrix dimensions of rows and columns of said data frame [11]. The code snippet is as simple as the one below.

```
df.shape
```

The output coming from this code is simply: (569, 13). What is interesting is that even though the official explanation of the attributes determines that the data set contains 12 attributes, the dimensions show 13. In fact, in the `head()` function output, there is a column at the end titled `Unnamed: 32` and is full of `NaN` instances. This column must have been an empty column stored as trash from the original table.

The final step for understanding the data set is to get a summary of the descriptive statistics of the columns, which is the purpose of function `describe()` [11]. The output of the said instruction is presented in Table 4.

Measure	diag	radius	texture	perimeter	area	smooth	compact	concavity	concave pts.	symmetry	fractal
count	569.00	569.00	569.00	569.00	569.00	569.00	569.00	569.00	569.00	569.00	5.69e+02
mean	0.37	14.13	19.29	91.97	654.89	0.10	0.10	0.09	0.05	0.18	6.28e-02
std	0.48	3.52	4.30	24.30	351.91	0.01	0.05	0.08	0.04	0.03	7.06e-03
min	0.00	6.98	9.71	43.79	143.50	0.05	0.02	0.00	0.00	0.11	5.00e-02
25%	0.00	11.70	16.17	75.17	420.30	0.09	0.06	0.03	0.02	0.16	5.77e-02
50%	0.00	13.37	18.84	86.24	551.10	0.10	0.09	0.06	0.03	0.18	6.15e-02
75%	1.00	15.78	21.80	104.10	782.70	0.11	0.13	0.13	0.07	0.20	6.61e-02
max	1.00	28.11	39.28	188.50	2501.00	0.16	0.35	0.43	0.20	0.30	9.74e-02

Table 4: Summary of the descriptive statistics of the data frame using the `describe()` function.

What Table 4 presents to us is a quantile and central tendency calculation of all columns in the data set. By looking at the 50% measure, which is the middle quantile or the *median* of each column, we can see that the center of the columns' distributions is quite similar in some attributes: radius and perimeter seem to have a median around the same level, and then smoothness, compactness, concavity and concave points also have a relatively close median but between the range of 0.00-0.10, whereas fractal dimension column appears to be the smallest measure. These differences in the medians mean that their distributions are **centered in different points along its x axis**, which can indicate the need of a **Standardization Transformation** in order to center all these distributions around a common range, which will get discussed later in more detail. Overall, the *mean* of a column tends to be bigger than the *median* of said column, which suggests **a positive skewness in the majority of the distributions**, and with this, **more outliers as part of the long right tails anticipated in the majority of the distributions**. If we are anticipating long right tails in the distributions, this can be confirmed when we look at 75% measure, or the upper quantile, and we can see that **the upper quantile (75%) tends to be more separated from the middle quantile (median) in comparison to the lower quantile (25%)**, which explains the the positive skewness and longer right tails.

## 1.2 Data Visualization: Further Understanding

At the end of the previous section, we discussed mainly the quartile calculations of each column, where we suggested that there are indicators of positive skewness along most of the attributes. However, all this assumptions where made from the analysis of numeric computations. When talking about distributions and its shape, the numeric central tendency measures do help, but it might be easier and visually confirmed by using plots.

In Figure 1, we made use of Seaborn's library function `distplot()`, which, as described in Seaborn's Documentation [13], is a combination of Matplotlib's `hist()` function that plots the frequency histogram of a numeric attribute, and a `kdeplot` that shows a distribution density estimate using a Gaussian Kernel.

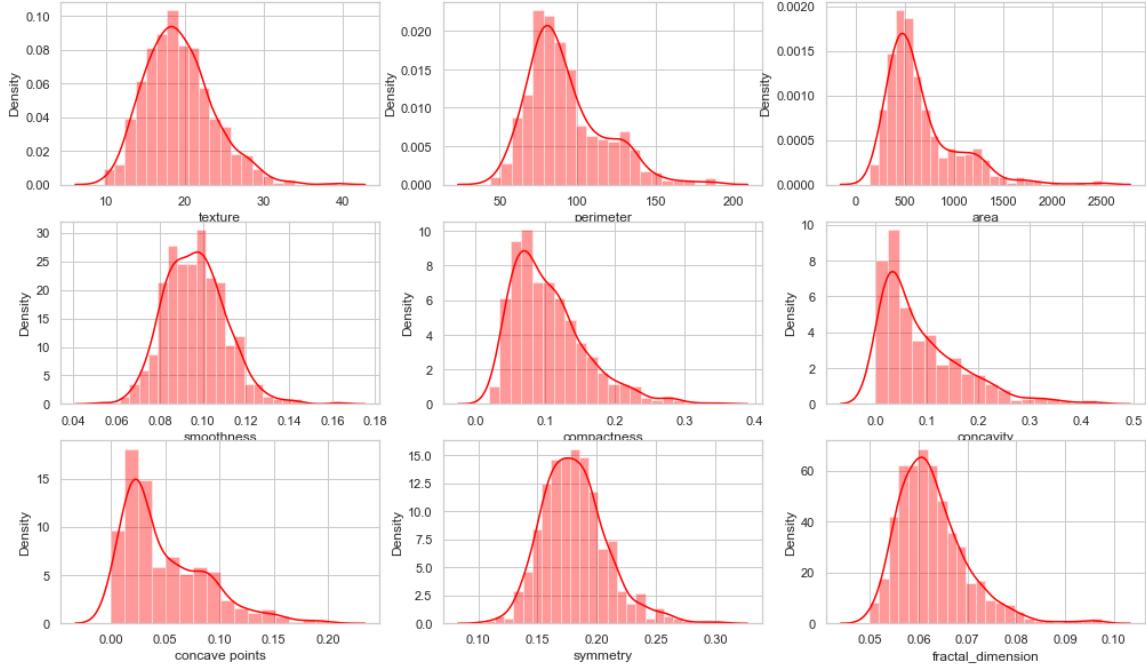


Figure 1: Distribution histogram and density plot of (almost) all attributes in the data set.

Thus, as implied by Figure 1, we confirm what we suggested at the end of the previous section, which

was the **Positive Skewness** present in almost all attributes when the distribution is, instead of symmetrically centered around the mode, concentrated towards the left side of the x axis, leaving a long tail in the left side. This tendency of concentration towards the left side moves the mode and the median towards this side as well, whereas the mean is carried towards the tail in the opposite side, which also explains why on Table 4 the mean almost always appeared larger than the median.

These conclusions invariably talk about where does the **majority of data concentration is**, which also describe the dispersion of the data found in each attribute. Now, conclusions about dispersion are more specifically visualized with a **box plot** of the data, and for this we will use the function `plot()` with parameter `kind='box'`, which is a Pandas function that plots a box plot per attribute of the data frame [11]. In Figure 2, the individual box plots of each attribute (left) and a common-axis box plot of all the plots together (right) is presented.

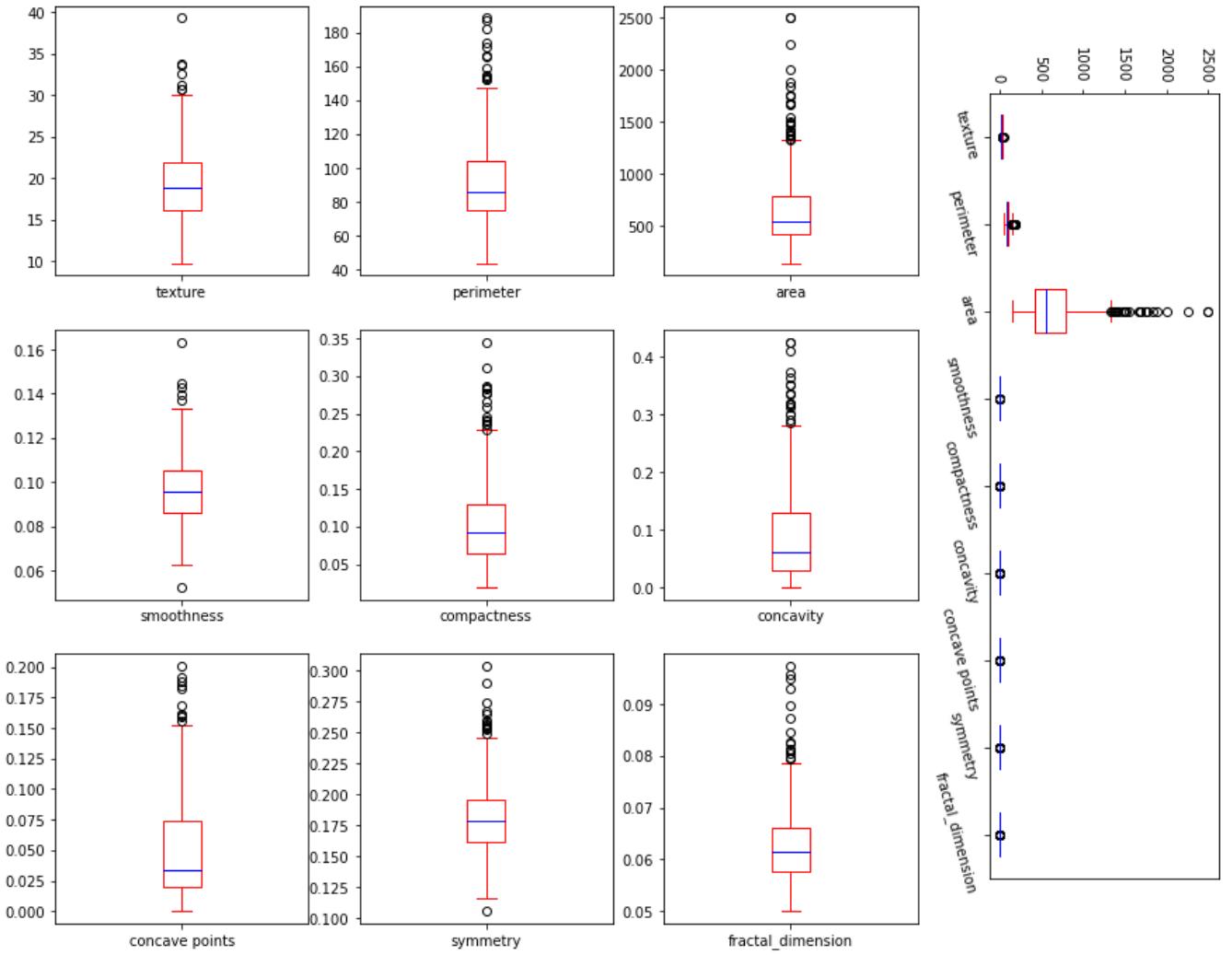


Figure 2: (Left) Box plot per attribute. (Right) All attributes' box plot in the same axis.

From Figure 2 we can see that **the concentration in the left side of the axis seen in the distribution density plots** in Figure 1 is now translated into a series of box plots that have **the box in the lower Y axis values, which represents this positive skewness** mentioned before. Also, in the right part of the figure a plot with all box plots is shown so that the dimensions of each attribute can be perceived visually: **the differences in the range of values of each attribute with one another is so large**

that most of the box plots are reduced to a line, showing specifically how the area column differs considerably from all other ranges. This will definitely need a Standardization Transformation so that their means and medians are centered near the same values and thus ease the process of comparison. It is also evident that **the amount of outliers is significant**, which explains why the *mean* of the majority of the attributes presented in Table 4 is dragged towards the right-side tail or the upper Y axis of the box plot.

Now that we have understood single-variable relationships in the data set, the next step is to understand the two-variable relationships, and for that a **correlation matrix** is computed. A **correlation matrix** is a matrix that computes the correlation coefficient, which is the covariance of two variables divided by their individual standard deviations [6], for all possible permutations of two attributes among the data set, and then displayed in a matrix form for visual purposes. This correlation matrix is computed using Pearson method as a parameter to Pandas function `corr()` and then that matrix is sent to Seaborn's `heatmap()` plot function, as shown in the code snippet below.

```
correlation = df.corr(method='pearson')
plt.figure(figsize=(10,5))
ax = sea.heatmap(correlation, annot = True, cmap='plasma', vmin=-1, vmax=1)
# some axis and label formatting code [...]
plt.show()
```

The Seaborn's `heatmap()` plot function receives the matrix of correlation coefficients in order to give each matrix cell a color according to the value that is in the cell [14], so that a better visualization of the matrix data can be done using scaled colors. Figure 3 shows such heatmap for all correlations of all pairs of attributes in the data set.

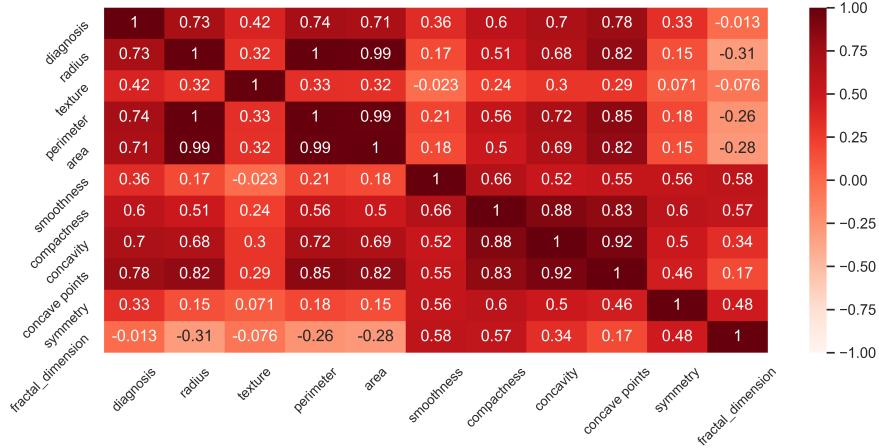


Figure 3: Correlation matrix for all the pairs of attributes in the data set.

If we take a look at the first row of Figure 3, we can see the correlation of all the attributes with the target column `diagnosis`, and it is noticeable that all but three attributes have a correlation bigger than 0.5, meaning that **the majority of attributes are strongly positively related to the target column**, which will increase the performance of the Machine Learning classifier method, since this high correlation means that the attributes can estimate or anticipate strongly the behaviour of the classification column. Also, we can notice that the correlations between the attributes is **particularly strong among the area, perimeter and radius columns**, due to the obvious relationship that this measures have in the computation of one another, where both the formulas for area and perimeter need the radius, and thus can be anticipated by knowing this measure.

We just mentioned the fact that most attributes have a **strong positive correlation** with the target column *diagnosis*, but what does this mean? well, the correlation that is positive tends to draw a straight line with positive slope, whereas the negative association draws a linear behaviour but with negative slope. Once again, the correlation matrix showed us the numeric computation, but in order to understand better this linear relationships, a plot helps in the visual perception of said correlations. For this, Seaborn has the plot function `pairplot()`, which plots pairwise relationships in a data set in the form of scatter plots for all possible pair permutations [15]. The output of this is shown in Figure 4 (the label names might look too small in the matrix, but they follow the same order as in the correlation matrix in Figure 3).

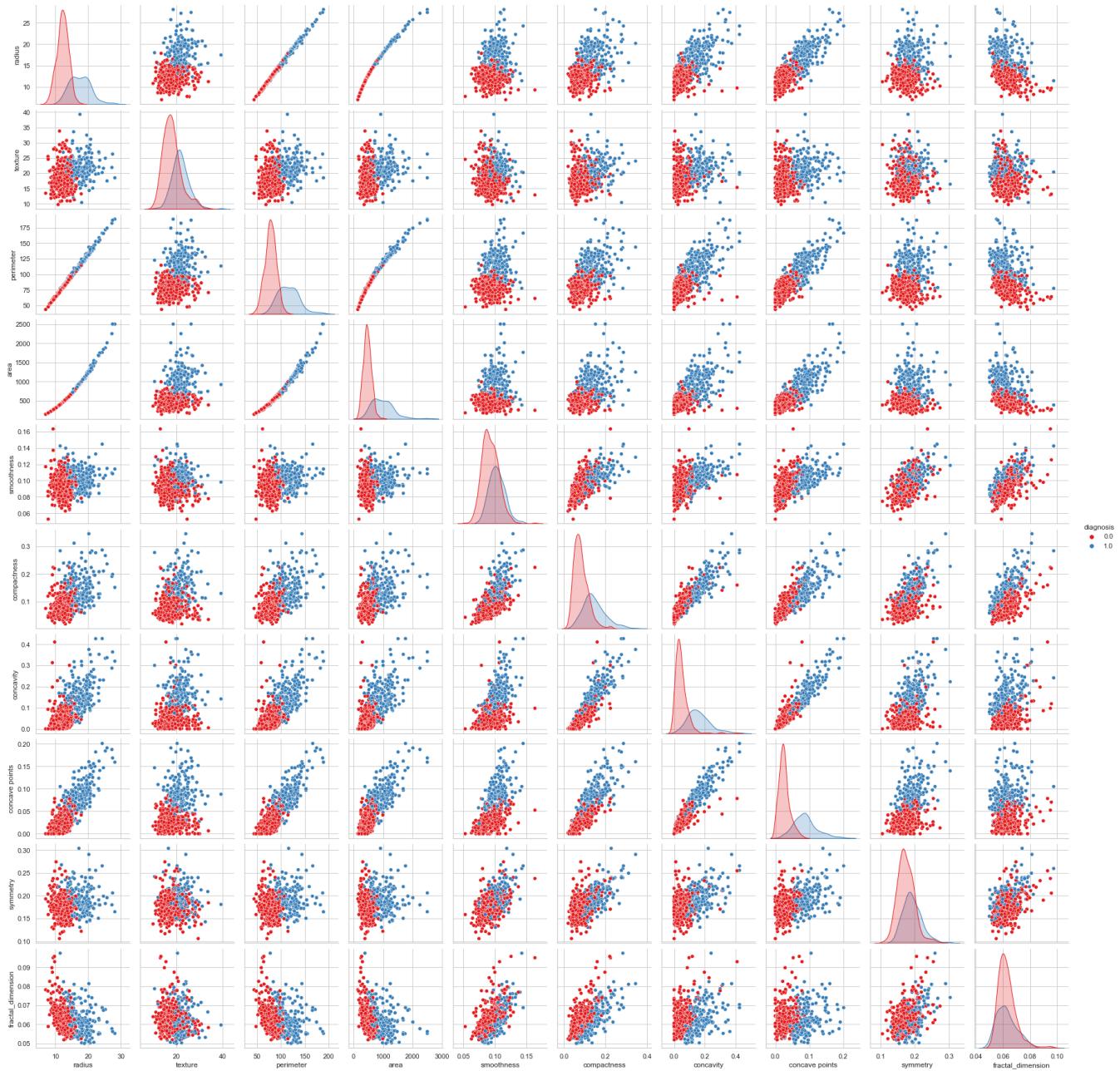


Figure 4: Scatter plots of all combinations of attribute pairs in matrix form.

From Figure 4, we can perfectly see that the tumor area, perimeter and radius almost follow a perfect straight line that goes upwards, describing what the correlation matrix told us about these columns with its high and positive value. This scatter matrix lets us see only the pair relationship within the attributes,

not with the target column. Nevertheless, we can see that there is always a positive tendency in the data: the cloud of points is somehow always pointing upwards, except for the fractal dimension and symmetry rows (last two), which show mostly data with no apparent trend. Excluding these two attributes, we can almost confidently say that also all points in blue (Malign) and red (Benign) are visually separable scatters. The overall conclusion from Figure 3 and Figure 4 is that there might be some **data redundancy** within the columns of area, perimeter and radius, with radius being the one with less interaction with the target column (lower correlation), and thus the most redundant. Also, the fact that the two classes are somehow visually separable along all cells in the scatter matrix will help in the performance of the classifier algorithm in further sections.

## 2 Data Set Cleaning

### 2.1 Erasing Unrelated Columns

For the purposes of training a Logistic Regression algorithm, we need to binarize the **diagnosis** column, so that instead of B and M we have 0 and 1 respectively as the possible values for this attribute, making it our class column. Additionally, the **id** attribute does not represent a numeric attribute per se, because its values do not intervene in the classification of a tumor. Thus, this column will better be dropped, as well as a column named **Unnamed: 32**, mentioned in the previous section, that unexpectedly appears at the end of all the mentioned columns, and is filled with **NaN** entries. In this way, we are making our data set to have **only numeric columns**, and **one binary class column** that determines the classification attribute, which makes our problem **supervised binary classification**. The following code snippet performs this dropping of columns operation using Pandas' **del** operator [11].

```
bcancer['diagnosis'] = bcancer['diagnosis'].apply(lambda x: 1 if x == 'M' else 0)
del bcancer['Unnamed: 32']
del bcancer['id']
```

### 2.2 Dropping Off Based On Correlation

Now, after having our data set as only numeric columns, the next step is to look for **data redundancy**. An attribute (column or feature) of a data set is considered **redundant** if it can be derived from any other attribute or set of attributes [8]. This can be perfectly identified whenever a column presents a strong correlation to another attribute, which usually means a correlation close to 1. For a quick visualization of the correlations in the data set, a heat map was plotted using the calculated correlation with the *Pearson* method for all the possible combinations of column pairs, presented in Figure 3.

Coming back to the correlation plot, indeed there are some strongly correlated attributes, which show values close to 1, such as: **perimeter**, **area** and **radius**. The attributes **perimeter** and **area** present a correlation of 0.99 because of their calculation involving other similar measures, but they are completely different measures of the main properties of the tumor, so they were not dropped from the data set. Thus, the only pair remaining with a strong correlation value is **radius** and **perimeter**. Which one to delete? The column **perimeter** presents a correlation of 0.74 with the **diagnosis** class column, whereas the column **radius** presents a correlation of 0.73 with the **diagnosis**, therefore, **radius** can be considered as redundant, and thus is dropped from the data set. In other words, we delete the **radius** attribute due to its redundancy and lower relationship with the class. This operation is shown in the snippet below.

```
del df['radius_mean']
df.head()
```

At the end of these operations we can observe some instances that are left in the data set, shown in Table 5.

class	texture	perimeter	area	smoothness	compactness	concavity	concave points	symmetry	fractal
1	10.38	122.80	1001.0	0.1184	0.2776	0.3001	0.1471	0.2419	0.07871
1	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	0.1812	0.05667
1	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	0.2069	0.05999
1	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	0.2597	0.09744
1	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	0.1809	0.05883

Table 5: Instances after cleaning operations (diagnosis is class).

Thus, our data set seems ready to be plugged into the learning models. It is also important to mention that after performing the following,

```
benign = df.groupby('diagnosis').size()[0]
malign = df.groupby('diagnosis').size()[1]
total = malign + benign
benign_p = benign / (total) * 100
malign_p = malign / (total) * 100
print(f"benign: {benign_p:.2f}%")
print(f"malign: {malign_p:.2f}%")
```

we can see that 62.74% of the instances correspond to class 0 (Benign) and 37.26% belongs to class 1 (Malign). These percentages, although different, can conclude that the data set is somehow balanced in terms of number of instances per class, to avoid an over-training of the model towards a specific class value, which would arise if the percentages differed much more.

In the following sections, we will perform 5 different resampling techniques that are used for the validation process: 1) Cross Validation, 2) Cross Validation with Repetitions, 3) Division by Percentage, 4) Division by Percentage with Repetitions and 5) Leave One Out Cross Validation. Additionally, each of these 5 techniques is performed 3 times: one with the raw data up until this point, another with the Standardized data set, and a last one with the data set transformed using Yeo-Johnson transformer. This is done with the purpose of, at the end of these operations, discussing the differences in both performance and estimated accuracy depending on the version of the data set that was used, and hopefully, conclude which transformation seems the one to proceed with to the Algorithm Evaluation section. After each of the 3 executions of a model adjustment, a measure called **accuracy** is presented, which is the average of the performance of each generated **sample**. This accuracy is a good estimate of the future performance of the data set when given new data to predict, and thus is used for the conclusions of each section.

### 3 Resampling Methods: Cross Validation

#### 3.1 Raw Data

The first Resampling Method is the Cross Validation, in which we split the data into N parts and dedicate 1 to test the model and the N-1 remaining to train it, and iterate N times, each time changing the test part until it has moved across the whole data set [9]. We apply this method to **raw data**, which means that the **x** variable contains a matrix of the data set values **without any transformation**, and the class column left out as **y** variable. The Logistic Regression model adjustment needed in this case around **3500 iterations** before it converged.

```

num_folds = 10
kfold = KFold(n_splits=num_folds, shuffle=True)
model = LogisticRegression(solver='lbfgs', max_iter=3500)
results = cross_val_score(model, x, y, cv=kfold)
acc = results.mean() * 100
stdev = results.std()
print(f"accuracy: {acc:.2f}% std: {stdev:.2f}")

```

As a result we get an **accuracy of 90.51%** and a standard deviation of **0.0335** for this method applied with raw data. Requiring 3500 iterations for the ML modeling seems risky, since this is a method that can potentially involve repetitions, and thus cost a lot more time, computationally speaking. With further methods this will show up more clearly as a serious issue.

### 3.2 Standardized Data

Now we will transform the **x** data matrix using a transformation called Standardization. This operation on matrix **x** involves that, for each column  $x_i$ , the following is applied to get every instance:

$$x_{scaled} = \frac{x_i - \min(x_i)}{\max(x_i) - \min(x_i)} \quad (1)$$

This transformation **translates** the attribute distribution towards a mean of around 0 and a standard deviation of 1 [5], which was constantly needed in previous sections because the ranges of the attributes differed so much that its comparison was difficult. If we plot the new standardized data columns, we can see such results in Figure 5. This transformation only translates the distribution to a more centered range, in simple words; this means that the distribution of the attributes is not modified towards a complete Gaussian bell, because its purpose is to standardize the ranges of values. This standardized data set will be called now **rescaled x**. We can now apply the model adjustment using **rescaled x**.

```

num_folds = 10
kfold = KFold(n_splits=num_folds, shuffle=True)
model = LogisticRegression(solver='lbfgs', max_iter=100)
results = cross_val_score(model, rescaledX, y, cv=kfold)
acc = results.mean() * 100
stdev = results.std()
print(f"accuracy: {acc:.2f}% std: {stdev:.2f}")

```

The Logistic Regression model adjustment done above outputs an **accuracy of 94.19%** and a standard deviation of **0.024** for the method applied to standardized data. Compared to the method in raw data, the accuracy in the prediction of the class by the model increased significantly from 90.51% to 94.19%, but more importantly, the iterations needed were reduced so that the model can converge in just 100 iterations, **reducing the computation time and increasing the accuracy**. This reduction of computation time is due to **reduction of the searching space** that the model needs to look for, meaning that the algorithm reaches the final model in less time because the data that it has to adjust to is closer to one another, whereas in raw data the ranges differ so much the algorithm takes more time to adjust its parameters.

As a side note, with the new standardized data, we can plot out box plots using one axis, because now the distributions of all attributes are centered around the same level. This is shown in the top plot in Figure 7.

### 3.3 Standardization and Yeo-Johnson Transformed Data

The **rescaled x** matrix of instances, as seen in Figure 5, is centered in a mean of zero and has a unit standard deviation, but it can suffer further transformations so that the column instances have a more

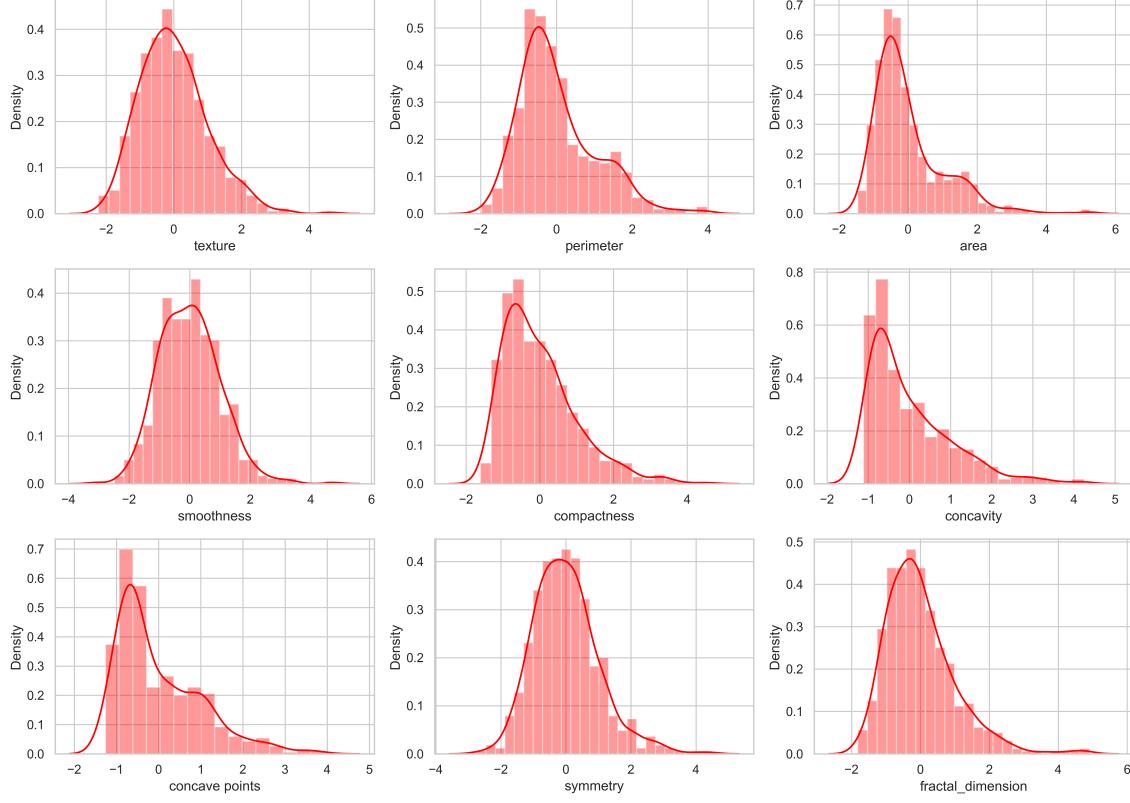


Figure 5: Rescaled  $x$ : Standardized data attributes.

Gaussian-like distribution, eliminating the strong positive skewness mentioned earlier. This transformation can be achieved by using either the *Box-cox method* or the *Yeo-Johnson method*, but in this case the *Yeo-Johnson method* is the only one of the two that accepts negative data values [4], such as the ones present in our data set. The method tries to fit the best parameters to achieve a Gaussian bell with each of the **rescaled  $x$**  attributes, producing an  $x$  matrix we will refer as **transformed  $x$** , plotted in Figure 6.

If we compare Figure 5 and Figure 6, we can see that not every column was transformed by Yeo-Johnson's method, because not all attributes seemed to have a significant enough skewness to be transformed, such is the case of `texture`, `smoothness` and `symmetry` columns. This was concluded after analysing which box plots presented a strong skewness towards some side of the other, such the top plot in Figure 7. As a result, the bottom plot in Figure 7 shows also how the **number of atypical instances was reduced by Yeo-Johnson's method**, which most of the times reduces the solution search process in a Machine Learning algorithm.

With **transformed  $x$**  matrix ready, we compute now the Cross Validation to this data in the code snippet below.

```

num_folds = 10
kfold = KFold(n_splits=num_folds, shuffle=True)
model = LogisticRegression(solver='lbfgs', max_iter=100)
results = cross_val_score(model, transformedX, y, cv=kfold)
acc = results.mean() * 100 # percentage of accuracy
stdev = results.std() * 100
print(f"accuracy: {acc:.2f}% std: {stdev:.2f}%)"

```

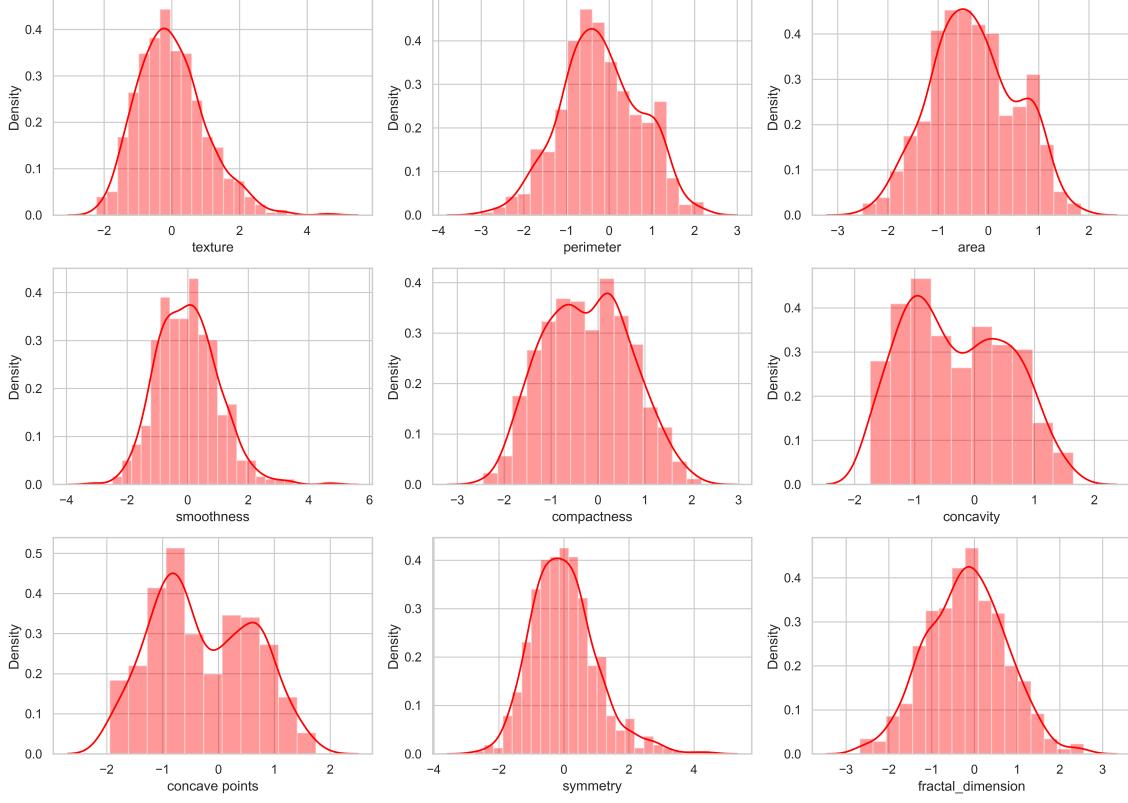


Figure 6: Transformed  $x$ : Standardized and Yeo-Johnson transformed data attributes.

Unfortunately, the **accuracy did not increase from last method, since now we get 91.56% with a stdev of 0.0271**, but it does increase with respect to the method applied in raw data. The algorithm also converges in around 100 iterations, so this method also involves a reduction from computation time when compared to raw data since the distributions are standardized by Yeo's transformer.

## 4 Resampling Methods: Cross Validation with Repetitions

### 4.1 Raw Data

Next up, the Cross Validation with Repetitions consists in, as its name implies, the performance of repeated Cross Validation methods. But, why would it help to repeat this method over and over? Because Cross Validation itself chooses the part that will be test **randomly**, and thus, by repeating it several times, we will reach further testing of our data with different combinations of  $N$  sets [9]. We go back to applying this method to the raw data matrix  $x$ , as in the code snippet below.

```

num_folds = 10
num_repeated = 5
repeatedkfold = RepeatedKFold(n_splits=num_folds, n_repeats=num_repeated)
model = LogisticRegression(solver='lbfgs', max_iter=3500)
results = cross_val_score(model, x, y, cv=repeatedkfold)
acc = results.mean() * 100 # percentage of accuracy
stdev = results.std()
print(f"accuracy: {acc:.2f}% std: {stdev:.2f}")

```

Did it help? Well, the **accuracy in prediction was raised from 90.51% to 90.65% and the standard deviation from 0.0335 to 0.0370 with respect to Cross Validation with raw data**,

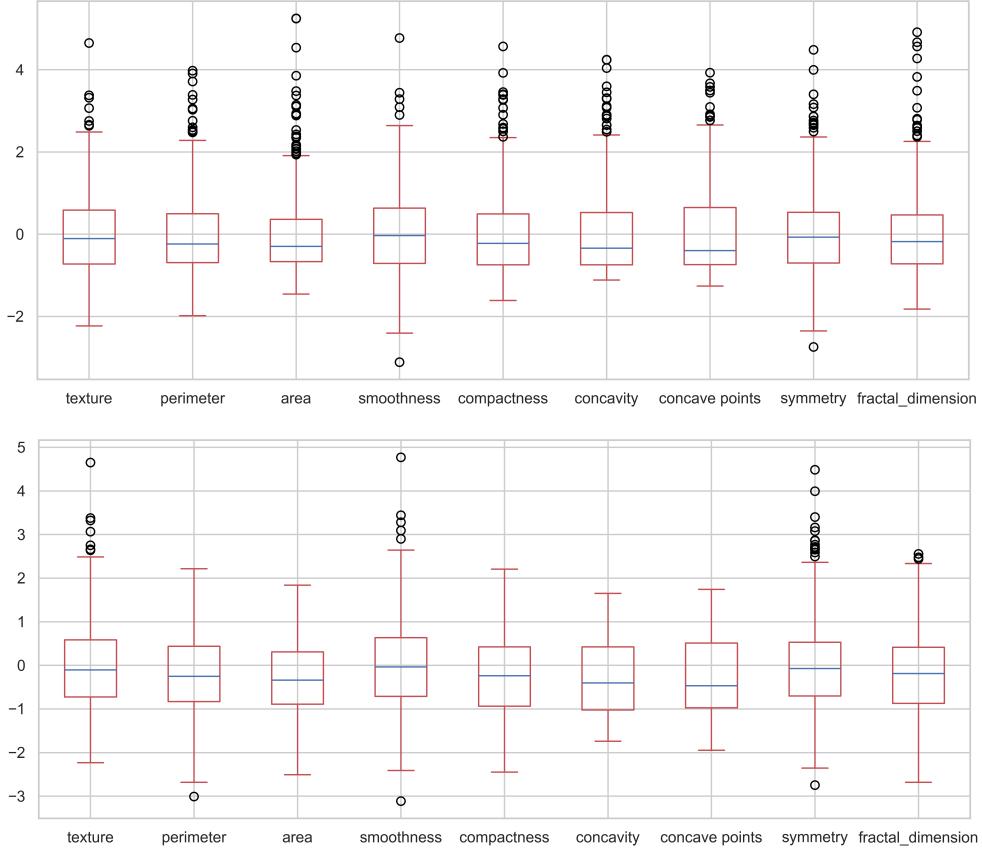


Figure 7: Rescaled x (top) Transformed x (bottom): box plot comparison.

which is somehow a growth in accuracy when raw data is used, but it is so small that it is likely to be insignificant and variate because of the randomness of the method. The fact that the standard deviation grew also a small bit is probably due to the fact that we simply have now more combinations of sets. Thus, at least with this data set, the repeated Cross Validation seems to be redundant and not significantly helpful when it comes to using raw data.

## 4.2 Standardized Data

Taking back the **rescaled x** data matrix, which is plotted in Figure 5, now it is time to see if the small increase in accuracy with raw data becomes a more significant increase in accuracy when applied to the standardized **x** matrix. We perform this Cross Validation with Repetitions below.

```

num_folds = 10
num_repeated = 5
repeatedkfold = RepeatedKFold(n_splits=num_folds, n_repeats=num_repeated)
model = LogisticRegression(solver='lbfgs', max_iter=100)
results = cross_val_score(model, rescaledX, y, cv=repeatedkfold)
acc = results.mean() * 100
stdev = results.std()
print(f"accuracy: {acc:.2f}% std: {stdev:.2f}")

```

The actual result was the decrease to 93.85% prediction accuracy and 0.0326 standard deviation decrease as well, and this decrease may mean the model without the repetitions was a bit

over-trained and now with more sets of testing and training, we reduced this 94% accuracy to a 93% due to the bigger amount of combinations.

### 4.3 Standardization and Yeo-Johnson Transformed Data

Now that we know what and why use Yeo-Johnson method, we can go back to the **transformed x** matrix, which includes standardization and Yeo-Johnson transform the the raw **x**.

```
num_folds = 10
num_repeated = 5
repeatedkfold = RepeatedKFold(n_splits=num_folds, n_repeats=num_repeated)
model = LogisticRegression(solver='lbfgs', max_iter=100)
results = cross_val_score(model, transformedX, y, cv=repeatedkfold)
acc = results.mean() * 100
stdev = results.std()
print(f"accuracy: {acc:.2f}% std: {stdev:.2f}")
```

Interestingly, the same thing occurred when we compared the Cross Validation with raw data to Cross Validation with Standard and Yeo's method, **we go from 90% to 91.60% accuracy**, while the highest (93.85%) accuracy was reached with only the Standard Transformation. The fact that the same behaviour from the simple Cross Validation and Cross Validation with Repetitions happens tells us that maybe **the iterative process is still redundant and gives the same results with or without the repetitions**.

## 5 Resampling Methods: Division by Percentage

### 5.1 Raw Data

Now we will validate our Logistic Regression model with another technique called Division by Percentage, which consists in splitting the data set **x** into two parts only, but how big are these parts? The two parts will be divided following a given percentage, often 33% for the training and the remaining 67% for the testing or validation [9]. This splitting is also random, but done once.

Let's use the raw data matrix **x** and apply this method to the trained model. Such code snippet is below.

```
test_size = 0.33
seed = 1
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size = test_size,
                                                    random_state = seed)
model = LogisticRegression(solver='lbfgs', max_iter=3500)
model.fit(x_train, y_train)
results = model.score(x_test, y_test)
acc = results * 100
print(f"accuracy: {acc:.2f}%")
```

The output in this case was **87.3% of prediction accuracy**, which is quite lower compared to the two methods seen before, with or without raw data. This is probably due to the method itself: it is just splitting and testing **once**, whereas the previous two methods involve inner iterations in the method itself. This method **lowers the accuracy level but also decreases the computation time and complexity**, allwing us to perform 3500 as usual with raw data, but in less than a second.

### 5.2 Standardized Data

Let us locate ourselves again with the standardized x matrix, called **rescaled x** and plotted in Figure 5, and try to apply this method to see how the model is validated: let's see the accuracy with **rescaled x**.

```

test_size = 0.33
seed = 1
x_train, x_test, y_train, y_test = train_test_split(rescaledX, y, test_size = test_size,
                                                    random_state = seed)
model = LogisticRegression(solver='lbfgs', max_iter=100)
model.fit(x_train, y_train)
results = model.score(x_test, y_test)
acc = results * 100
print(f"accuracy: {acc:.2f}%")

```

The results here are very worth noticing. Compared to the method with raw data, this method with standardized data **increases the prediction accuracy from 87.3% to 93.1%, which is the biggest percentage increase so far**. With a reduced computation time such as the one needed for a Division by Percentage Validation, we can achieve similar accuracy results to the ones by Cross Validation and Cross Validation Repeated, which may be a crucial conclusion if our model needs large amounts of data and the computation time is an important issue.

### 5.3 Standardization and Yeo-Johnson Transformed Data

Changing our data back to **transformed x**, which is transformed by Standardization and Yeo-Johnson's Method, let's take a look at what the validation accuracy is with the code snippet below.

```

x_train, x_test, y_train, y_test = train_test_split(transformedX, y, test_size =
                                                    test_size, random_state = seed)
model = LogisticRegression(solver='lbfgs', max_iter=100)
model.fit(x_train, y_train)
results = model.score(x_test, y_test)
acc = results * 100
print(f"accuracy: {acc:.2f}%")

```

Once again, we see that by taking **transformed x** as the input data, **the prediction accuracy increases (91.49%), but not as much as with the standardized data**. This seems to be quite interesting, since the Yeo-Johnson transforms the set to be more similar to a Gauss bell, which is the data that a Machine Learning model such as Logistic Regression expects and works best with. This certainly will be addressed in the final conclusions of this document, as it seems to happen in all methods.

## 6 Resampling Methods: Division by Percentage with Repetitions

### 6.1 Raw Data

As mentioned above, the Division by Percentage method operates only one split and test over the input data, but takes the 33% randomly. Now, taking advantage of that randomness, there exists another variation of this method called Division by Percentage with Repetitions. This method validates our model by taking 33% for testing and 67% of our data as training, but several times, since the split is done randomly [9]. This method looks similar to what a simple Cross Validation does, because it makes N iterations to shuffle the part taken as test around our data set. The code snippet below shows the implementation of a Division by Percentage with Repetitions method with raw data matrix **x**.

```

test_size = 0.33
n_splits = 10
kfold = ShuffleSplit(n_splits=n_splits, test_size = test_size)
model = LogisticRegression(solver='lbfgs', max_iter=3500)
results = cross_val_score(model, x, y, cv=kfold)
acc = results.mean() * 100
stdev = results.std()

```

```
print(f"accuracy: {acc:.2f}% std: {stdev:.2f}")
```

With raw data, we get **a prediction accuracy of 91.38% and a std deviation of 0.016**, which means that if we perform the same method of Division by Percentage over and over (10 times in this case), the accuracy of the model with respect to the correct and known  $y$  increases from **87% to around 91%** using raw data, which is the same 91% we get from Division by Percentage with **transformed  $x$** , but by doing so simply with raw data  $x$ . Thus, we can say that Division by Percentage with Repetitions applied to raw data has the same output as a Simple Division by Percentage done with Standardized and Yeo-Johnson transformations.

## 6.2 Standardized Data

Taking back the Standardized matrix **rescaled  $x$** , we can apply the Repeated Division by Percentage:

```
test_size = 0.33
n_splits = 10
kfold = ShuffleSplit(n_splits=n_splits, test_size = test_size)
model = LogisticRegression(solver='lbfgs', max_iter=100)
results = cross_val_score(model, rescaledX, y, cv=kfold)
acc = results.mean() * 100
stdev = results.std()
print(f"accuracy: {acc:.2f}% std: {stdev:.2f}")
```

Similarly, using the standard matrix we get **93.67% of prediction accuracy and a standard deviation of 0.0094**, which is the smallest deviation we have got in any previous method, showing that the accuracy results in each repetition do not vary a lot throughout the 10 iterations. This can mean that the repetitions may be unnecessary, which is somehow shown also in the accuracy itself, because from raw data and standardized data we only increased around 2%, whereas other methods applied to standard data show a bigger increase.

## 6.3 Standardization and Yeo-Johnson Transformed Data

We can apply this method of validation now with **transformed  $x$** , which includes Standard and Yeo-Johnson transforms, to see if redundancy is also suggested:

```
test_size = 0.33
n_splits = 10
kfold = ShuffleSplit(n_splits=n_splits, test_size = test_size)
model = LogisticRegression(solver='lbfgs', max_iter=100)
results = cross_val_score(model, transformedX, y, cv=kfold)
acc = results.mean() * 100
stdev = results.std()
print(f"accuracy: {acc:.2f}% std: {stdev:.2f}")
```

Interestingly, we got **accuracy of 91.97% and standard deviation of 0.0018**, which is almost the same accuracy result we got with this same method but with raw data, but slightly bigger this time. This is a strange behaviour, since we were noticing that with **transformed  $x$**  the accuracy in prediction grew a bit more than just 0.59%. This might show that this method done over and over is not really being helped in accuracy by the transformations, but these just reduce the computing time from 3500 iteration to 100 iterations in the model convergence itself.

## 7 Resampling Methods: Leave One Out Cross Validation

### 7.1 Raw Data

Finally we have reached the last validation method in the scope of this document: Leave One Out Cross Validation. What this means is, in iteration 1, what it does is take just one column as test data and the remaining columns as training data, and so on. Depending on the number of columns (N), the number of iterations there will be. We basically take all data values as training except one, instead of a percentage, which goes to test [9]. The standard deviation will be higher, because we will be comparing one data value against its predicted value for an error estimation. Let's see how this performs with raw data  $\mathbf{x}$ :

```
loocv = LeaveOneOut()
model = LogisticRegression(solver='lbfgs', max_iter=3500)
results = cross_val_score(model, x, y, cv=loocv)
acc = results.mean() * 100
stdev = results.std()
print(f"accuracy: {acc:.2f}% std: {stdev:.2f}")
```

Surprisingly, this method took over 12 seconds to output the result, which is quite noticeable compared to all the other methods that take at best 1 second to output the calculations. Either way, we get **accuracy in prediction of 90.69% and a standard deviation of 0.29** which is similar to Cross Validation, Cross Validation Repeated and Division by Percentage Repeated, but done in much more time, which might look as a disadvantage of this method, but quite logical due to the amount of iterations depending on the data set columns.

### 7.2 Standardized Data

Taking back the Standardized matrix **rescaled  $\mathbf{x}$**  in Figure 5, we can apply the Leave One Out CV:

```
loocv = LeaveOneOut()
model = LogisticRegression(solver='lbfgs', max_iter=100)
results = cross_val_score(model, rescaledX, y, cv=loocv)
acc = results.mean() * 100
stdev = results.std()
print(f"accuracy: {acc:.2f}% std: {stdev:.2f}")
```

Not surprisingly, the **accuracy increased 94.20% and the standard deviation decreased to 0.23**, which is quite the same increase from raw to standard data we have seen in all the previous methods. But now, the computation time is not so large as this time it took around 2 seconds, which confirms in a more exaggerated way that **the standardization of data reduces computation time since it makes the Logistic Regression converge faster in less iterations**.

### 7.3 Standardization and Yeo-Johnson Transformed Data

Changing our data back to **transformed  $\mathbf{x}$** , which is transformed by Standardization and Yeo-Johnson's Method, let's take a look at what the validation accuracy is with the code snippet below and the Leave One out CV method.

```
loocv = LeaveOneOut()
model = LogisticRegression(solver='lbfgs', max_iter=100)
results = cross_val_score(model, transformedX, y, cv=loocv)
acc = results.mean() * 100
stdev = results.std()
print(f"accuracy: {acc:.2f}% std: {stdev:.2f}")
```

This result was almost exactly the same as the one with raw data  $\mathbf{x}$ , where now the **accuracy** is **91.74%** with a standard deviation of **0.27**, but slightly increased and done also in 100 iterations (2 seconds), compared to the 3500 iterations that raw data required. Thus, the process done using **rescaled  $\mathbf{x}$**  shows a bigger accuracy percentage, which has been happening with all methods.

## 8 Resampling Methods: Conclusions

All the accuracy percentages we discussed might look at this point confusing and hard to synthesize, that is why all the accuracy percentages throughout the different methods are plotted in Figure 8.

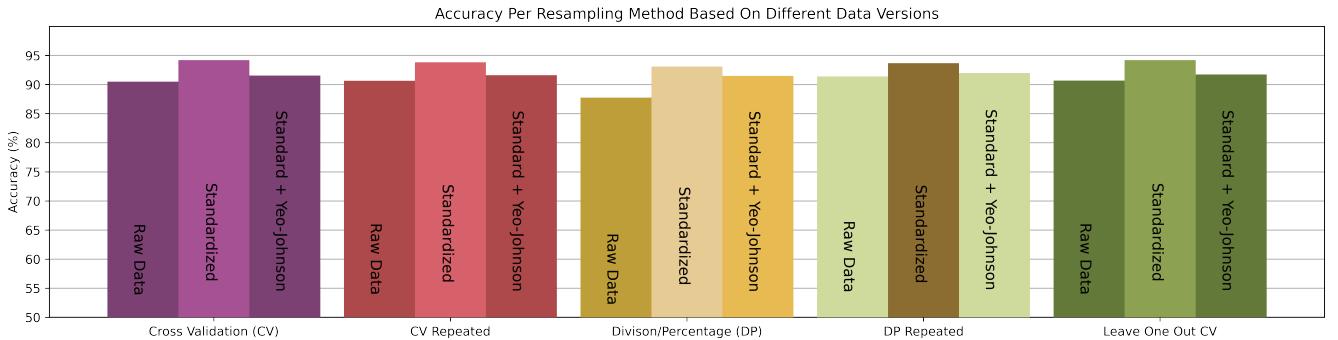


Figure 8: Prediction accuracy mean results by validation method.

Now we can easily conclude that, throughout every validation method, the lowest accuracy was reached when using raw data, followed by when the method used Standardized and Yeo-Johnson transformed data, and the highest accuracy was reached when using only Standardized data. The fastest but lowest accuracy can be reached with Division by Percentage method, while the slowest method was Leave One Out Cross Validation, due to the iterations depending on the amount of columns of the data set.

Nevertheless, the accuracy in prediction using all these methods is quite high in general, surrounding the values of 87%-94%. This might look as too high even, and most probably will appear as an over-trained model when new input data is provided outside of the provided database and the accuracy decreases. I could be mistaken however, and the new input data set might as well result in similar accuracy levels, meaning that the model was not over-trained but actually extremely precise, which could be the case since Kaggle's data sets are very curated and maintained. And if this is true, the reason behind this model's precision (high accuracy) might rely on what we see in Figure 3: **almost all attributes present a correlation coefficient higher or equal to 0.5** with respect to the target column, meaning that the attributes in this data base are **strongly positively linearly correlated, resulting in a Logistic Regression model that shows high levels of accuracy**, since the attributes' values strongly indicate a consistent change of the class column with a positive trend.

The most notorious outcome of this method comparison seen in Figure 8 is how the **Standardization of raw data increased significantly the prediction accuracy of a model while reducing the convergence time**, compared to the model adjusted to raw data. This is due to the fact that raw data has **attribute ranges that differ greatly from other attributes'**, which **increases the searching space for the Machine Leraning model applied**, making the iterative method take longer to converge to the optimized solution. After Standardization of data, **the ranges of all attributes are similar, and thus the searching space is reduced, which makes the iterative method converge faster and all attributes have an equal influence on the model calculation**.

However, one thing that indeed surprises is the accuracy level with a Standard and Yeo-Johnson transformation: the purely Standardized data set decreases the computation time in all methods compared to the method applied with raw data, but **with a Standard and Yeo-Johnson transformation the accuracy is increased compared to raw data but decreased compared to only Standardized data**. This is surprising because, in theory, Machine Learning algorithms such as Logistic Regression actually expect data centered in  $\mu = 0$  and  $\sigma = 1$  (Standardized) and that resembles a Gaussian Bell distribution, which is exactly what Yeo-Johnson's method does to data. But why is the accuracy not the highest with this two transforms that deliver data as expected by Logistic Regression? Let's look at Figure 9.

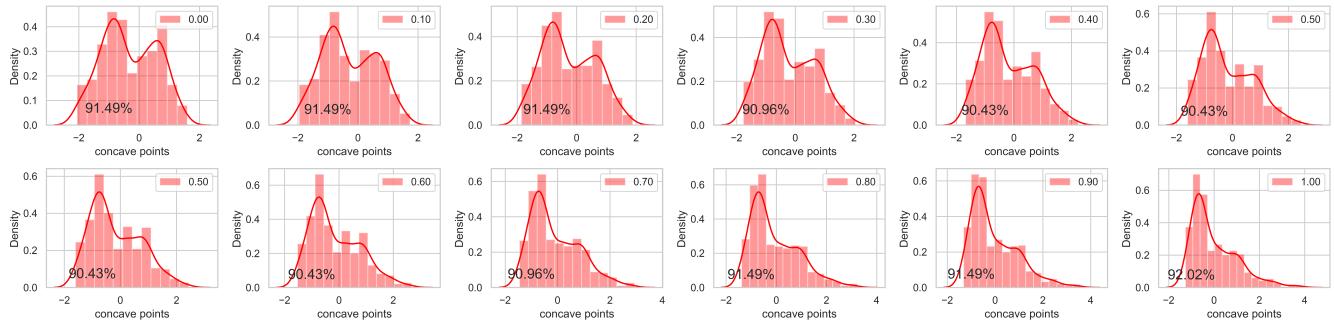


Figure 9: Lambda custom setting for Yeo-Johnson Transformation applied to one column.

The answer to this might rely on the *lambda* values that Yeo-Johnson's method fits to each column. By looking at Figure 9, where a sample column, in this case `concave points`, is taken to suffer a custom lambda setting each time in order to cover the range from 0 to 1 and then calculate the new accuracy of the model with this transform, we can see that there is one value of this lambda that increase the accuracy percentage (percentage shown in the lower left corner of each plot in Figure 9) of the **overall** set of data columns, which is exactly a lambda of 1.00, and which also represents a no transform value that leaves the column with the same distribution shape as in Figure 5. If we set this column to a transformation with a lambda of 1.0, we increase the overall accuracy to 92.02%, whereas the fitted lambda value of the method (0.11) shows an overall accuracy of 91.49%. Thus, the conclusion might be that Yeo-Johnson method can increase the accuracy of the model, but some manual changes to the calculated lambdas are needed.

Overall, the conclusion for this data set specifically would be that the fastest method is Division by Percentage, with the data version being the Standardized data set, so that the model can converge in a much lesser amount of iterations. Cross Validation seems to do a better job at the controlled but still random sets of data, producing the highest accuracy with Standardized data, but takes slightly more time. For the Algorithm Evaluation section, the transformation and resampling technique that seems both fast and relatively accurate for our problem is **Standardized Data Set applied to a Division By Percentage resampling technique**.

## 9 Algorithm Evaluation: Metrics

The **accuracy** we have been presenting as a metric for the resampling methods only tells us if the final prediction is the same as the expected (known) output of a Binary Classification Problem. There are other metrics that tell us how good is our algorithm with respect to each of the classes, in a separate way. This is helpful when for example we have a database that has 90% of the entries belonging to class 0 and only 10% belonging to class 1: this metric will tell you how good is the algorithm for the prediction of each class. In that case, we would notice that the percentage of accuracy for the prediction of class 0 is very high compared to the accuracy for class 1, for example, and there we would notice a fault in our database:

unbalance. The current data set has 63% of instances classified as Benign tumors and the remaining 37% of the instances are classified as Malign tumors, and while this is still inside the category of *balanced*, there is a majority of data classified as one class, and thus the following metrics will tell us separately how the classifier works per class [10].

Now that the data set has been transformed with a Standardization and resampled with Division by Percentage, we can continue with the actual metrics that evaluate the performance of a Machine Learning Model, which in this case is a Logistic Regression Model. The scope of this document is to use two metrics: Cohen's Kappa Score, and AUC (Area Under the Curve) for the ROC Curve. In order to understand the said metrics, a concept is presented: Confusion Matrix. Thus, it is important to reaffirm that we will continue from now on using **rescaled x**, which is the Standardized data set, since the last section's conclusion was that the time performance was reduced and the resampling that showed a modest accuracy was Division by Percentage. There are other transformations such as MinMax or an L2 Normalization, which normalize columns based either on an specific range or based on Euclidean Distance [2], but according to the data in the attributes, the measures' purposes would be affected and thus the data set would no longer be related to predicting tumors.

## 9.1 The Confusion Matrix

The **Confusion Matrix** is the first computation that gives us an idea of the precision per class that was just mentioned. The Confusion Matrix is a calculation that shows how much concordance can there be between two or more points of view or classes [10], in this case only two. The matrix will have a size of  $n \times n$ , where  $n$  is the number of classes, thus our matrix is of size  $2 \times 2$ . What goes into each of the four cells is presented in Figure 10,

		Prediction	
		1	0
Real	1	True Positives TP	False Negative FN
	0	False Positive FP	True Negatives TN

Figure 10: The Confusion Matrix cells' content.

where the concepts are defined as follows:

- True Positive (TP): where both Real and Predicted coincide for 1's (Positives).
- True Negative (TN): where both Real and Predicted coincide for 0's (Negatives).
- False Positive (FP): Prediction (False) says Positive (1) and Real says Negative (0).
- False Negative (FN): Prediction (False) says Negative (0) and Real says Positive (1) [10].

Now, let's see how our data set outputs the Confusion Matrix. For that purpose, the following code snippet was used. As said before, we will continue from now on using **rescaled x**, which is the Standardized data set, since the conclusion was that the time performance was reduced and the resampling that showed a modest accuracy was Division by Percentage (see Section 8).

```
from sklearn.metrics import confusion_matrix
test_size = 0.33
seed = 1
x_train, x_test, y_train, y_test = train_test_split(rescaledX, y, test_size = test_size,
                                                    random_state = seed)
model = LogisticRegression(solver='lbfgs', max_iter=1000)
model.fit(x_train, y_train)
predicted = model.predict(x_test)
matrix = confusion_matrix(y_test, predicted, labels=[1,0]) # labels param so that the
# classes coincide with the indices
```

The confusion matrix that results from the above computation and accommodated in their corresponding cells is shown in Table 6.

		Predicted →	
		1	0
Real ↓	1	TP = 57	FN = 8
	0	FP = 5	TN = 118

Table 6: Confusion Matrix of the current data set.

We can interpret from the matrix that the sum of all cells is 188, instead of 569. This is due to the fact that we are evaluating the resampling technique Division by Percentage, which uses a **test size of 33%** as if it were a real-life testing sample whose output we are evaluating with the Confusion Matrix. Thus, the 33% of 569 rows is  $187.77 \approx 188$ . In this way, we see that the True Negatives (TN) are 118 and the True Positives (TP) are 57. This has a logical explanation because the TN (0,0) = 118, or when both the predicted and real outputs coincide on 0 (which is our Benign tumor), is the class from which we have 63% of instances. Thus, the TP (1,1) = 57, or when both predicted and real outputs coincide on 1 (which is our number for Malign Tumor), is also the class for the remaining 37% of instances. What it is showing is that 118 and 57 have somehow a relationship that resembles the 63%-37% that the instances have, though 57 is actually around 48% of 118: this tells us **that most (67%, since  $118 + 57 = 175$ , and 118 is 67% of 175) of the right (according to the known data) predictions are True Negatives, and a test that is right (according to the known data) has a probability of around 33% of being True Positive ( $118 + 57 = 175$ , and 57 is 33% of 175)**. The False Positives and False Negatives maintain a similar relationship as well.

## 9.2 Cohen's Kappa Score

After the Confusion Matrix analysis, we can move forward to our next evaluation metric: **Cohen's Kappa Score**, which is an agreement index that oscillates in the range of  $[-1, 1]$  and is calculated as:

$$\kappa = \frac{P_o - P_e}{1 - P_e}, \quad (2)$$

where  $P_o$  is the Proportion/Probability of observed agreement, or what we presented in previous sections as **accuracy**, and where  $P_e$  is the Proportion of expected agreement by mere randomness [3]. A common interpretation would be that  $\kappa = 1$  represents perfect agreement and  $\kappa \leq 1$  represent no agreement. The more balanced the classes are, the higher the kappa score index will be. If we perform this operation on our **rescaled x** applied to the selected method Division by Percentage,

```

from sklearn.metrics import cohen_kappa_score
test_size = 0.33
seed = 1
x_train, x_test, y_train, y_test = train_test_split(rescaledX, y, test_size = test_size,
                                                    random_state = seed)
model = LogisticRegression(solver='lbfgs', max_iter=1000)
model.fit(x_train, y_train)
predicted = model.predict(x_test)
cohen_score = cohen_kappa_score(y_test, predicted)

```

we get as result **Cohen's Kappa Score = 0.84**, which is fewer than the **accuracy** presented to rescaled x in Section 5 (Division by Percentage) that was 93.1% or 0.931. This reduction is due to the fact that kappa score takes on account separate classes' prediction and also random guesses, not just the comparison of right and false predictions. Anyway, a kappa score of 0.84 is considered almost perfect agreement, and this confirms that our data set is quite balanced, which is triggering a kappa score close to 1.

### 9.3 ROC Curve and AUC Metric

The **ROC Curve** is a curve that moves along an X and Y axis, where X axis represents the False Positive Rate (FPR) and the Y axis is the True Positive Rate (TPR) calculated from the Confusion Matrix [1] in Table 6 as:

$$Sensitivity = TPR = \frac{TP}{TP + FN} = \boxed{0.8769}, \quad (3)$$

$$Specificity = TNR = \frac{TN}{TN + FP} = \boxed{0.9593}, \quad (4)$$

and thus,

$$FPR = 1 - Specificity = 1 - \frac{TN}{TN + FP} = \frac{FP}{TN + FP} = \boxed{0.0406} \quad (5)$$

Therefore, we have a coordinate point inside our ROC Curve axes, which is (FPR, TPR) or (0.0406, 0.8769). In this way, the ROC Curve is plotted fully by changing all the real values according to a gradually changing threshold, and each time the threshold moves through the data set rows, a new point (FPR, TPR) is plotted to form the ROC Curve [1]. Having this in mind, let's plot our ROC Curve for the **rescaled x** data set in the code snippet below,

```

from sklearn.metrics import roc_curve
from sklearn.metrics import RocCurveDisplay
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression

seed=1
x_train, x_test, y_train, y_test = train_test_split(rescaledX, y, test_size = test_size,
                                                    random_state = seed)
model = LogisticRegression(solver='lbfgs', max_iter=1000)
model.fit(x_train, y_train)
predicted = model.predict(x_test)
y_score = model.decision_function(x_test)
fpr, tpr, _ = roc_curve(y_test, y_score, pos_label=model.classes_[1])
roc_display=RocCurveDisplay(fpr=fpr, tpr=tpr).plot(color='blue')
plt.scatter([0.0406], [0.8769], marker='o', color='red', zorder = 10, s = 60)
plt.plot([0.0, 1.0], [1.0, 0.0], color='gray', zorder= -10, linestyle='dashed')

```

whose output is the middle plot in Figure 11.

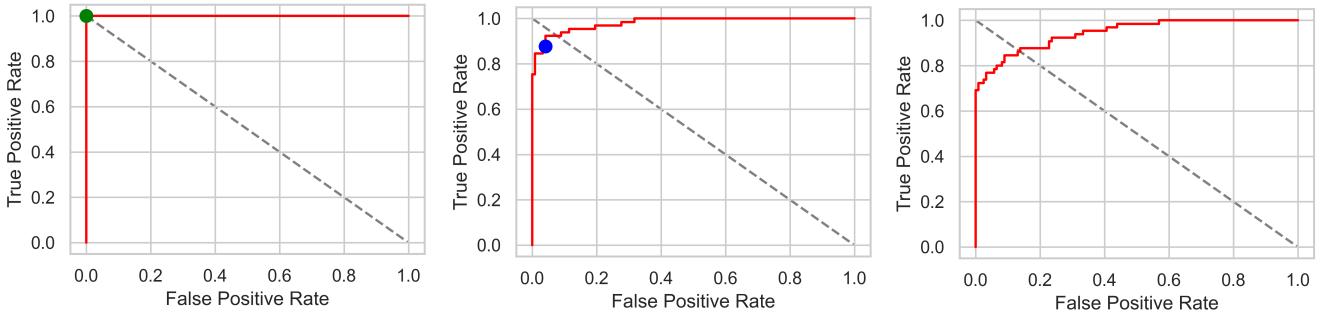


Figure 11: (Left) Ideal ROC Curve. (Middle) Rescaled x ROC Curve. (Right) Raw Data x ROC Curve.

The left plot in Figure 11 is the *ideal* ROC Curve for a Logistic Regression classifier model, where it is able to perfectly distinguish between all the Positive and Negative classes. The current standardized data set ROC Curve would be the middle plot in Figure 11, and the right plot is the ROC Curve if we performed the calculation in raw data  $x$ . We can see from the middle and right plot that the middle is closer to the ideal, but how to measure it exactly? The **Area Under the Curve (AUC)** of the ROC plot is the *measure of the ability of a classifier to distinguish* between classes and is used as a summary of the ROC curve [1]. After computing the AUC we get: for the left plot,  $AUC = 1$ ; for the middle,  $AUC = 0.98$ ; and for the left, the  $AUC = 0.96$ .

With a concrete measure for how good is the ROC Curve we can make our final conclusions: the standardization not only helps in the time performance when training a Machine Learning model like Logistic Regression, but also the AUC is improves, which means that the ability of the model to classify correctly is improved when compared to the AUC of the data without any trasformation. However, in all plots of Figure 11 there is a gray dashed line, which determines the *ideal* point where the first point (FPR, TPR) should have fallen over the plot if the data set was perfectly balanced: if our data set's blue point would have fallen in the intersection of the dashed line and the ROC curve (green point), that would have meant our data set has the same amount of entries for the class 0 and class 1. Therefore, since our blue point falls towards the left side of the dashed line, that means **our model had more negative entries than positives**, which leads to conclude that **the classifier is more able to classify data if it belongs to the negative class, or the Benign tumor class (class 0)**.

In any case, the point is slightly tilted towards the left side, but the AUC is still quite high and the point is rather very close to the ideal, meaning that the data set is balanced, and the standardization transformation improved our time performance, the accuracy and the classifier ability to predict overall. Maybe Yeo-Johnson with a very manual configuration of the lambda would have increased the performance, but for now this metrics show the signs of a great classifier.

## References

- [1] Aniruddha Bhandari. *AUC-ROC Curve in Machine Learning Clearly Explained*. URL: <https://www.analyticsvidhya.com/blog/2020/06/auc-roc-curve-machine-learning/>.
- [2] Scikit-learn Developers. *Preprocessing Data*. URL: <https://scikit-learn.org/stable/modules/preprocessing.html>.

- [3] Scikit-learn Developers. *sklearn.metrics.cohenkappa\_score*. URL: [https://scikit-learn.org/stable/modules/generated/sklearn.metrics.cohen\\_kappa\\_score.html](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.cohen_kappa_score.html).
- [4] Scikit-learn Developers. *sklearn.preprocessing.PowerTransformer*. URL: <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.PowerTransformer.html>.
- [5] Scikit-learn Developers. *sklearn.preprocessing.StandardScaler*. URL: <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html>.
- [6] Adam Hayes. *Correlation*. URL: <https://www.investopedia.com/terms/c/correlation.asp>.
- [7] UCI Machine Learning. *Breast Cancer Wisconsin (Diagnostic) Data Set*. URL: <https://www.kaggle.com/uciml/breast-cancer-wisconsin-data>.
- [8] José Octavio Camarena Méndez. *Data Redundancy*. URL: <https://github.com/the-other-mariana/data-mining/tree/master/week1-2>.
- [9] José Octavio Camarena Méndez. *Resampling Techniques. Lecture Notes*. URL: <https://github.com/the-other-mariana/data-mining/tree/master/week7>.
- [10] Sarang Narkhede. *Understanding Confusion Matrix*. URL: <https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62>.
- [11] The Pandas Development Team. *Pandas Documentation*. URL: <https://pandas.pydata.org/docs/>.
- [12] The Pandas Development Team. *pandas.DataFrame.groupby*. URL: <https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.groupby.html>.
- [13] Michael Waskom. *seaborn.distplot*. URL: <https://seaborn.pydata.org/generated/seaborn.distplot.html>.
- [14] Michael Waskom. *seaborn.heatmap*. URL: <https://seaborn.pydata.org/generated/seaborn.heatmap.html?highlight=heatmap#seaborn.heatmap>.
- [15] Michael Waskom. *seaborn.pairplot*. URL: <https://seaborn.pydata.org/generated/seaborn.pairplot.html>.