Applications of Data Science Methods on Breast Cancer Dataset

Elshan Gadimov

[1] Eötvös Loránd University, Budapest, Pázmány Péter stny. 1/C, 1117
[2] elshanqadimov7@gmail.com

Abstract. Breast cancer is considered one of the deadliest diseases. Only in the USA,1 out of 8 women develops breast cancer over the course of their lifetime. And for women in the US, its death rate is second only to lung cancer.
In this report, I will illustrate the results of classification, clustering, and frequent pattern mining techniques on the breast cancer data set. The breast cancer domain that was used for this report was obtained from the University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia. Thanks go to M. Zwitter and M. Soklic for providing the data.

Keywords: Classification · Clustering · Frequent Pattern Mining.

1 Introduction

The applications of Data Science is vast and one of them is to predict if the breast cancer is recurrent or not. The goal of this report is to share the results of classification and clustering methods as well as to figure out how frequent pattern mining techniques could be utilized to enhance the prediction. But first I will start with dataset information and the preprocessing steps that I took before the classification and modelling.

2 Dataset Information

2.1 Exploration

The given dataset which is a Breast Cancer dataset is consist of 286 rows and 10 columns. 9 of those columns are categorical. Only one feature called **deg-malig** (degree of malignancy) is ordinal numerical and ranges from 1 to 3. The column named **Class** is a crucial feature for the classification part. I will use it as a label and try to predict it on the test.

One value is missing from the **breast-quad** column and 8 values are missing from the **node-caps** column.

I have done visualization using the matplotlib library to get some insights from the data. Mostly I used histograms and countplots.

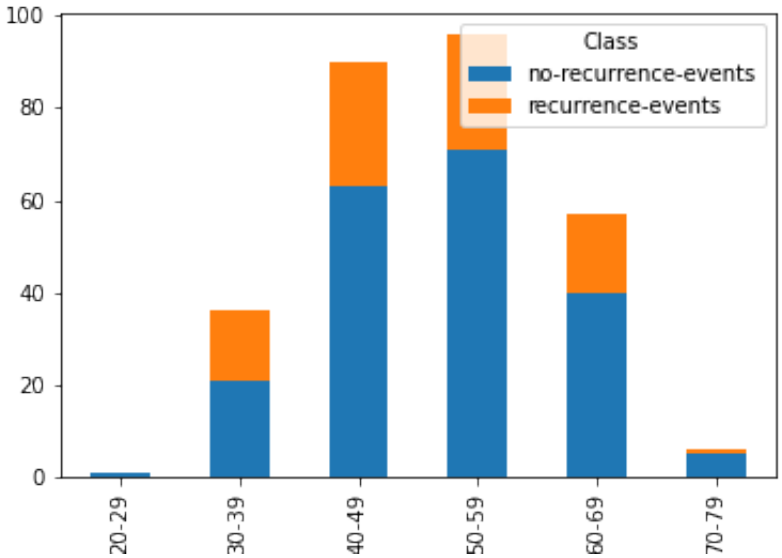

Fig. 1. The distribution of age intervals

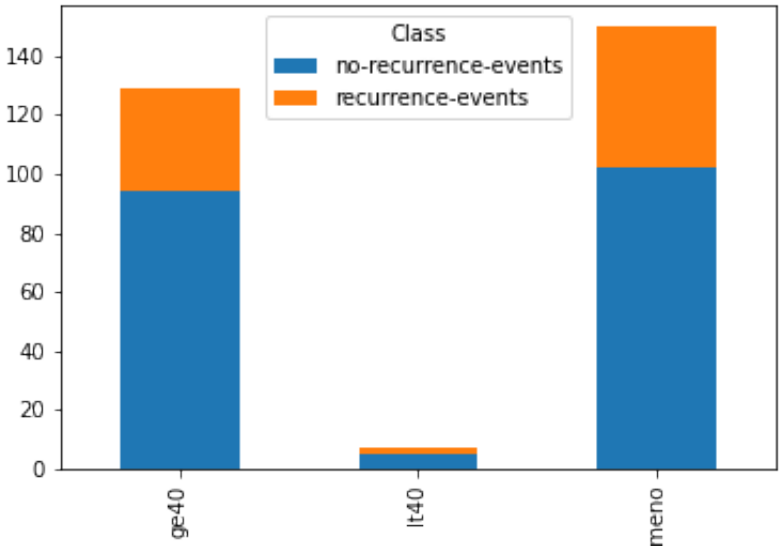

Fig. 2. Bar plot of menopause stacked with Class

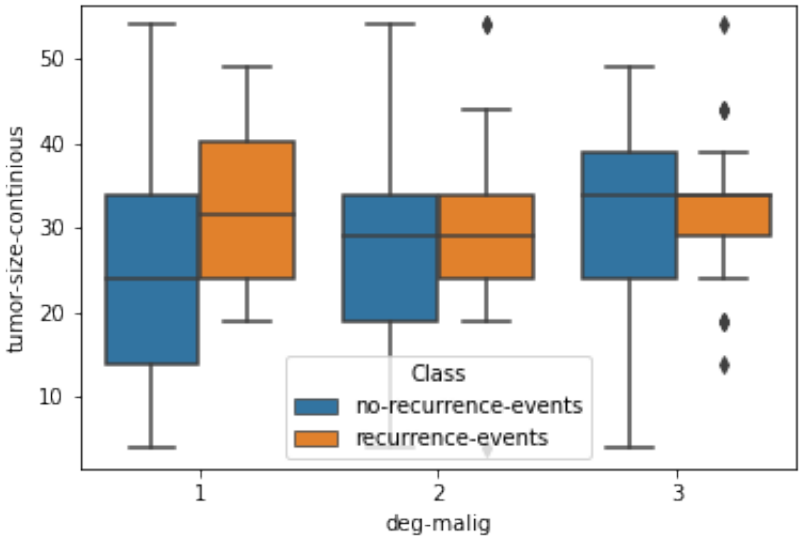

Fig. 3. Degree of Malignancy relation with tumor size

Based on plot shown in the **Figure 1** it is evident that most of the patients are in their 40s and 50s. Moreover, majority of recurrence cases fall into **premeno** and **ge40** cases of **menopause** column.

When I converted **tumor-size** which consisted of intervals to continuous and box plotted its relation with degree of malignancy there was an interesting result. When the degree of malignancy is equal 1 it is more likely for the cancer to be recurrent when the tumor size is big.

2.2 Data Prepossessing

Data Prepossessing is an important step as datasets are inconsistent and noisy. The low quality data will lead to poor data mining results. There are plethora of data preprocessing techniques used. For example, in data cleaning duplicates are removed, missing values are either dropped or imputed. I can not use data integration as I do not have access to the data warehouse.

Another step that was applied was One Hot encoding. One Hot encoding is applied to categorical features and it creates binary columns as the same amount of unique values on that column. Scaling was applied to the numerical features.

3 Classification

Classification is the problem of predicting the right label for a given input record. The task differs from regression in that labels are discrete entities, not continuous

function values. Trying to pick the right answer from two possibilities might seem easier than forecasting open-ended quantities, but it is also a lot easier to get dinged for being wrong. In our case I will apply classification algorithms to check if the the breast cancer recurrence events will happen or not. Many classification algorithms have been proposed by the researches over the years. Our objective is getting accuracy as much as possible. I will use 75 percentage of the data for the training and the remaining 25 percentage for the test. For this dataset we used algorithms like Logistic Regression, Random Forest Classifier and K-Nearest Neighbors.

3.1 Logistic Regression

Logistic regression is another technique borrowed by machine learning from the field of statistics. It is the go-to method for binary classification problems (problems with two class values). There are types of Logistic Regression. The first one is Binary and that is what I have applied on this dataset. The second is Multinomial and it is used when the target column has more than two different values. The last one is called Ordinal Logistic Regression and it used for rating. For instance, the service rating of a restaurant from 1 to 3. Another important step I have applied on the data was cross validation. It is a technique that is used to know how model performs on different parts of the dataset. Below are the results of 5 k folded model.

```
array([0.67241379, 0.71929825, 0.61403509, 0.71929825, 0.68421053])
```

The accuracy of the Logistic Regression model was 71% and the model was better at predicting **non-recurrent-events** than **recurrent-events**.

	precision	recall	f1-score	support
no-recurrence-events	0.73	0.87	0.79	46
recurrence-events	0.65	0.42	0.51	26
accuracy			0.71	72
macro avg	0.69	0.65	0.65	72
weighted avg	0.70	0.71	0.69	72

Fig. 4. Classification report of Logistic Regression

3.2 KNN

The k-nearest neighbors (KNN) algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems. The important factor in KNN is choosing right number for K.

	precision	recall	f1-score	support
0	0.75	0.92	0.82	51
1	0.56	0.24	0.33	21
accuracy			0.72	72
macro avg	0.65	0.58	0.58	72
weighted avg	0.69	0.72	0.68	72

Fig. 5. Classification report of KNN model

KNN model performed best when we set n neighbors to 4 and the accuracy was 74% on the test data.

3.3 Random Forest

Random forest is a class of ensemble methods specifically designed for decision tree classifiers. It combines the predictions made by multiple decision trees, where each tree is generated based on the values of an independent set of random vectors. The fundamental concept behind random forest is a simple but powerful one — the wisdom of crowds. In data science speak, the reason that the random forest model works so well is: A large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models.

Feature importance refers to techniques that assign a score to input features based on how useful they are at predicting a target variable.

The results of Random Forest Model were much better than the Logistic Regression model. Overall, accuracy of the model was 75% and like in the Logistic Regression model the precision of **non-recurrence-events** are still higher than **recurrence-events** and that is mostly due to the amount of **non-recurrence-events**

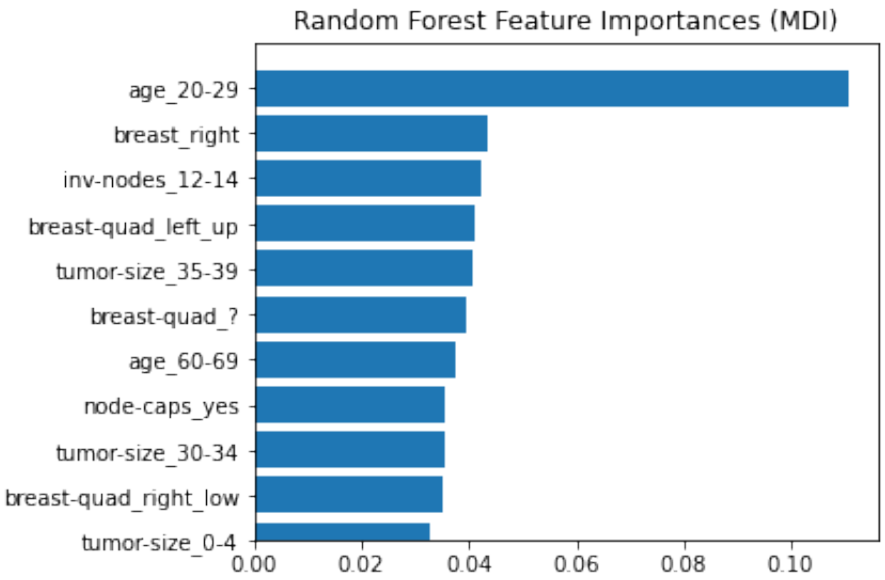

Fig. 6. Feature Importance done on the Random Forest Model

	precision	recall	f1-score	support
0	0.77	0.92	0.84	51
1	0.64	0.33	0.44	21
accuracy			0.75	72
macro avg	0.70	0.63	0.64	72
weighted avg	0.73	0.75	0.72	72

4 Clustering

Clustering is the problem of grouping points by similarity. Often items come from a small number of logical sources or explanations, and clustering is a good way to reveal these origins.

4.1 KModes

KMode is one of the unsupervised Machine Learning algorithms that is used to cluster categorical variables.KMeans uses mathematical measures (distance) to cluster continuous data. The lesser the distance, the more similar our data points are. Centroids are updated by Means. But for categorical data points, we cannot calculate the distance. So we go for KModes algorithm. It uses the dissimilarities(total mismatches) between the data points. The lesser the dissimilarities the more similar our data points are. It uses Modes instead of means.

I used the elbow method to find the right amount of the clusters. From the **Fig. 7.** we can conclude that 2 is the right amount of the clusters for this dataset.

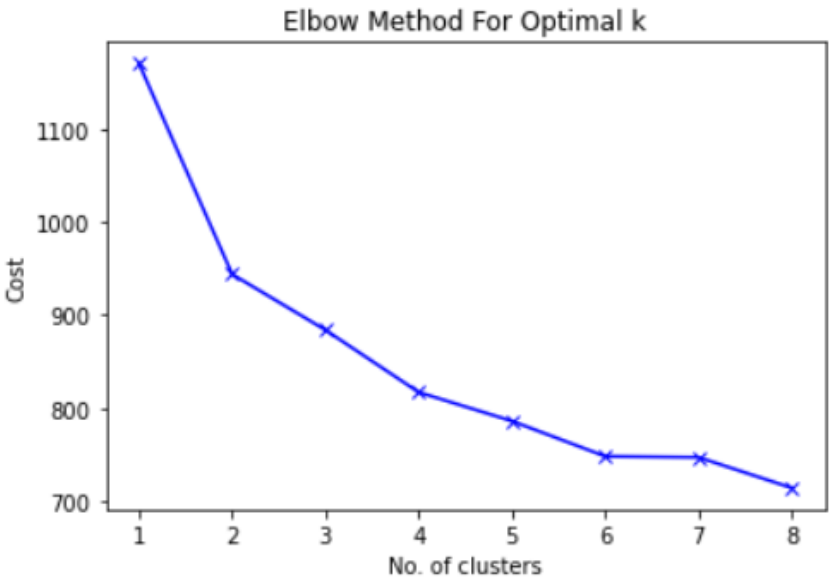

Fig. 7. Elbow method graph with KMode algorithm

4.2 Distance Based Clustering

Usually distance-based clustering algorithms can handle the categorical data but it is important to choose an appropriate distance function such as Gower's distance that combines the attributes as desired into a single distance. Later we can run clustering methods like Hierarchical Clustering, DBSCAN, OPTICS, etc. But still the results will never be **sound** with categorical data and our dataset very small consisting of only 286 rows.

$$D_{Gower}(x_1, x_2) = 1 - \left(\frac{1}{p} \sum_{j=1}^{p} s_j(x_1, x_2) \right)$$

Fig. 8. Gower Distance Formula

I decided to use Gower's distance and is computed as the average of partial dissimilarities across individuals. For the categorical values it is calculated as when values are equal the distance is 0 and whenever they are not equal they are calculated with the formula in **Fig. 8.**

```
KMeans(n_clusters=2)
Silhoutter Score 0.335
Davies Boildin Score 1.2941
Calinski Harabasz 119.0927


Birch(n_clusters=2)
Silhoutter Score 0.3039
Davies Boildin Score 1.438
Calinski Harabasz 111.6199


SpectralClustering(n_clusters=2)
Silhoutter Score 0.2875
Davies Boildin Score 1.224
Calinski Harabasz 59.0401
```

Fig. 9. Silhouette, Davies-Bouldin and Calinski-Harabasz Scores

I have used three clustering algorithms KMeans, Birch and Spectral Clustering and in **Fig. 9.** I calculated their results using Silhouette, Davies-Bouldin and Calinski-Harabasz scores. Silhouette score is better when it gets closes to 1 and bad when it approaches to -1. Near 0 silhouette score means clusters are overlapping. From the three algorithms I have tested KMeans performed better results than other two based on Silhouette and Calinski-Harabasz scores.

5 Frequent Pattern Mining

Clustering is the problem of grouping points by similarity. Often items come from a small number of logical sources or explanations, and clustering is a good way to reveal these origins.

5.1 KModes

KMode is one of the unsupervised Machine Learning algorithms that is used to cluster categorical variables.KMeans uses mathematical measures (distance) to cluster continuous data. The lesser the distance, the more similar our data points are. Centroids are updated by Means. But for categorical data points, we cannot calculate the distance. So we go for KModes algorithm. It uses the dissimilarities(total mismatches) between the data points. The lesser the dissimilarities the more similar our data points are. It uses Modes instead of means

6 First Section

6.1 A Subsection Sample

Please note that the first paragraph of a section or subsection is not indented. The first paragraph that follows a table, figure, equation etc. does not need an indent, either.

Subsequent paragraphs, however, are indented.

Sample Heading (Third Level) Only two levels of headings should be numbered. Lower level headings remain unnumbered; they are formatted as run-in headings.

Sample Heading (Fourth Level) The contribution should contain no more than four levels of headings. Table 1 gives a summary of all heading levels.

Table 1. Table captions should be placed above the tables.

Heading level	Example	Font size and style
Title (centered)	**Lecture Notes**	14 point, bold
1st-level heading	**1 Introduction**	12 point, bold
2nd-level heading	**2.1 Printing Area**	10 point, bold
3rd-level heading	**Run-in Heading in Bold.** Text follows	10 point, bold
4th-level heading	*Lowest Level Heading.* Text follows	10 point, italic

Displayed equations are centered and set on a separate line.

$$x + y = z \tag{1}$$

Please try to avoid rasterized images for line-art diagrams and schemas. Whenever possible, use vector graphics instead (see Fig. 1).

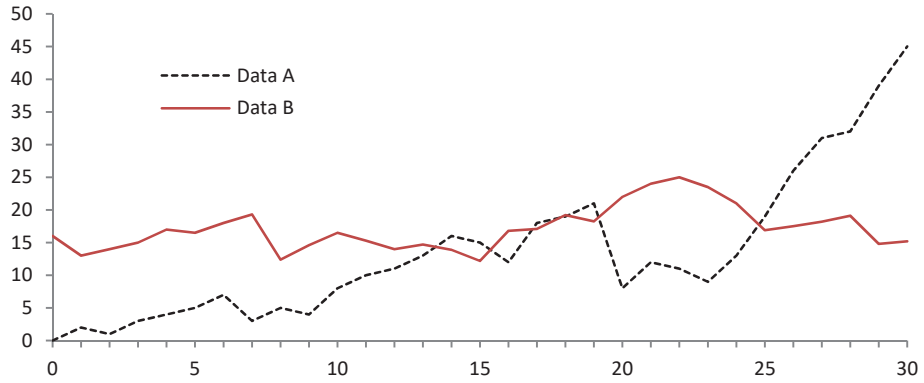

Fig. 10. A figure caption is always placed below the illustration. Please note that short captions are centered, while long ones are justified by the macro package automatically.

Theorem 1. *This is a sample theorem. The run-in heading is set in bold, while the following text appears in italics. Definitions, lemmas, propositions, and corollaries are styled the same way.*

Proof. Proofs, examples, and remarks have the initial word in italics, while the following text appears in normal font.

For citations of references, we prefer the use of square brackets and consecutive numbers. Citations using labels or the author/year convention are also acceptable. The following bibliography provides a sample reference list with entries for journal articles [1], an LNCS chapter [2], a book [3], proceedings without editors [4], and a homepage [5]. Multiple citations are grouped [1–3], [1, 3–5].

References

1. Author, F.: Article title. Journal **2**(5), 99–110 (2016)
2. Author, F., Author, S.: Title of a proceedings paper. In: Editor, F., Editor, S. (eds.) CONFERENCE 2016, LNCS, vol. 9999, pp. 1–13. Springer, Heidelberg (2016). https://doi.org/10.10007/1234567890
3. Author, F., Author, S., Author, T.: Book title. 2nd edn. Publisher, Location (1999)
4. Author, A.-B.: Contribution title. In: 9th International Proceedings on Proceedings, pp. 1–2. Publisher, Location (2010)
5. LNCS Homepage, http://www.springer.com/lncs. Last accessed 4 Oct 2017