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 in 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.

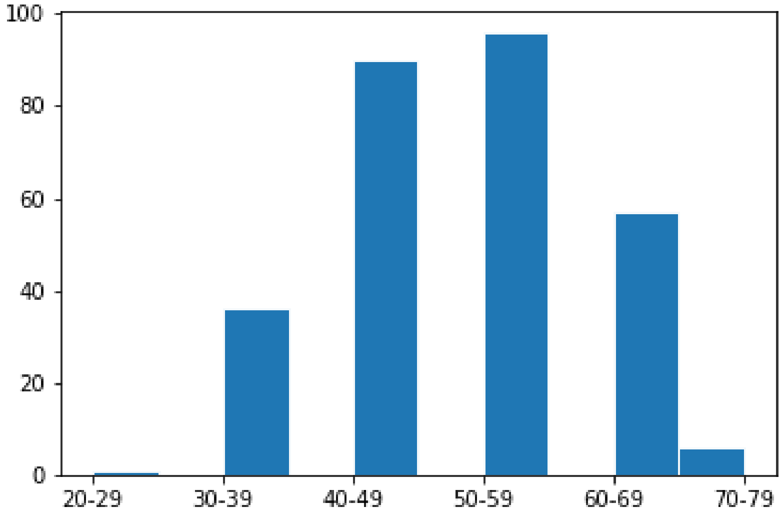

Fig. 1. The distribution of age intervals

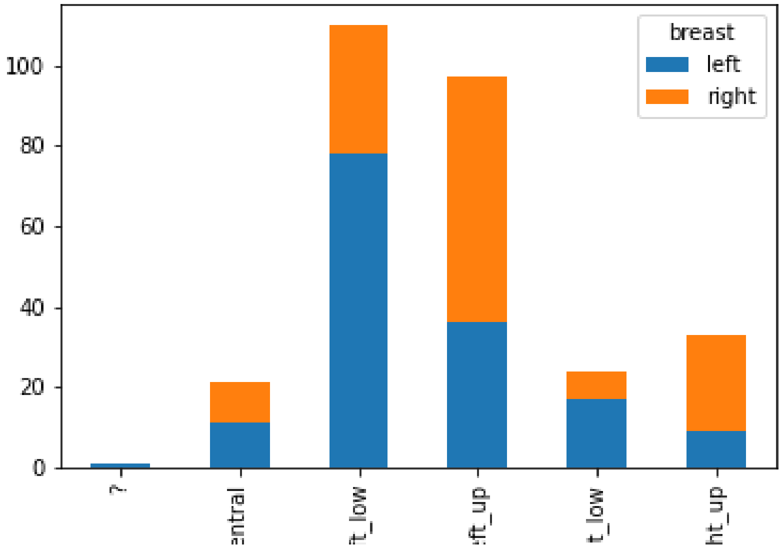

Fig. 2. Bar plot of breast quads stacked by Class

3 First Section

3.1 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. 2).

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].

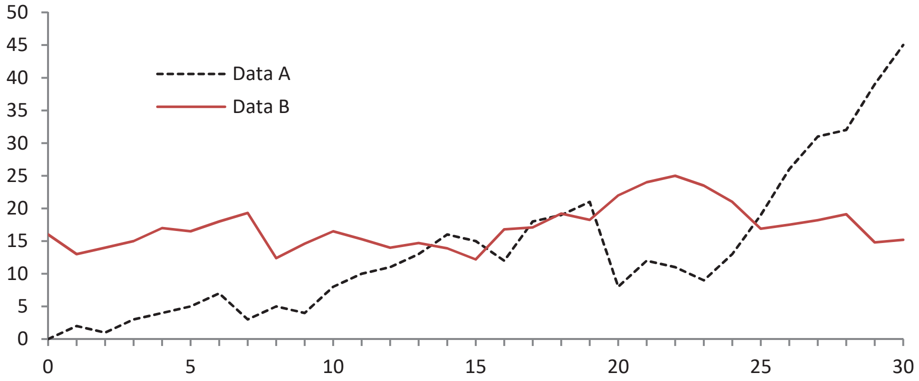

Fig. 3. A figure caption is always placed below the illustration. Please note that short captions are centered, while long ones are justified by the macro package automatically.

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