Introduction to Data Science, Semester Project

Mohammad Damerji[1]

[1]MSc. Artificial Intelligence, ELTE University - Faculty of Informatics

Abstract. This report introduces many supervised and unsupervised learning techniques used with the breast cancer data set, which belong back to the University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia. The project task is to build models to classify the data into no-recurrence-events or recurrence-events ones, cluster them into various groups, and uncover some patterns to get some useful information. In order to do that, after exploring, preprocessing, finding the characteristics of the data, the Gaussian Naive Bayes, a Random Forest Classifier, KNeighbors Classifier and KMeans algorithm, and some frequent pattern approaches are used.

Keywords: Breast cancer. Classification. Clustering. Frequent Pattern Mining.

1 Introduction

Cancer represents the fifth leading cause of death worldwide. In 2015, 1.7 million people died as revealed by the World Health Organization. According to this report, breast cancer is the most common among women in both developed and developing countries and represents 16 % of female cancer. Breast cancer survival rates vary widely across the world, from 80 % or more in North America, Sweden and Japan, to around 60 % in middle-income countries, to below 40 % in low-income countries. [1] [2]

To achieve the report's goal, python will be used with these libraries: NumPy for linear algebra and fixing arrays, Pandas for data processing, Seaborn for cool visualizations, matplotlib for figures, and Scikit-Learn for preprocessing and machine learning algorithms.

2 Exploration and Preprocessing

2.1 Load and explore data

As 9 Attributes in the dataset, so the task is to explore the data type, check if there any nan values, clean, visualize, encode the data, and

try to reduce the number of features in the dataset by deleting the not contributing ones, which can lead to:

- Accuracy improvements
- Overfitting risk reduction
- Speed up in training
- Improved Data Visualization

The dataset consists of 286 instances with nine attributes and one column as a label, it contains both categorical and numerical variables. See Fig. 1.

	Class	age	menopause	tumor-size	inv-nodes	node-caps	deg-malig	breast	breast-quad	irradiat
0	no-recurrence-events	30-39	premeno	30-34	0-2	no	3	left	left_low	no
1	no-recurrence-events	40-49	premeno	20-24	0-2	no	2	right	right_up	no
2	no-recurrence-events	40-49	premeno	20-24	0-2	no	2	left	left_low	no
3	no-recurrence-events	60-69	ge40	15-19	0-2	no	2	right	left_up	no
4	no-recurrence-events	40-49	premeno	0-4	0-2	no	2	right	right_low	no

Fig. 1. Dataset with heads and Categorical variables

The distribution of the classes (no-recurrence-events / recurrence-events) in the dataset are shown in Fig. 2.

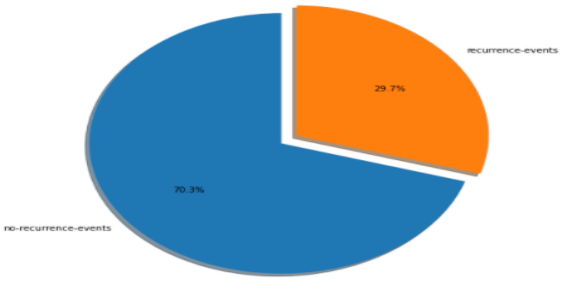

Fig. 2. The distribution of classes

The above Pie chart shows that about 70.3 % of the instances are considered as “no-recurrence-events” ones, where 29.7 % of the instances are considered as “recurrence-events”.

2.2 Handling the missing values

From the data description, we know that there are some missing values, they are in “node-caps”, “breast-quad” attributes.

```
There are missing values denoted by (?) in these attributs ['node-caps', 'breast-quad']
```

Fig. 3. The missing values

“node-caps” represents the evidence that cancer cells so it does not make sense to predict values instead of the missing values and the same for “breast-quad” attribute so we deleted all instances which have missing values. The remaining data consists of 277 instances,

now, we have about 70.8 % of the instances are considered as “no-recurrence-events” ones, where 29.2 % of the instances are considered as “recurrence-events”.

2.3 Visualize and pre-processing the data

By plotting the “breast” attribute, we found that the number of instances which belong to “no-recurrence-events” class is approximately equal to the number of instances which belong to “recurrence-events” class in the two cases, in other words, this attribute should be deleted because it does not matter where the cancer is located on the right or left side. See Fig. 4.

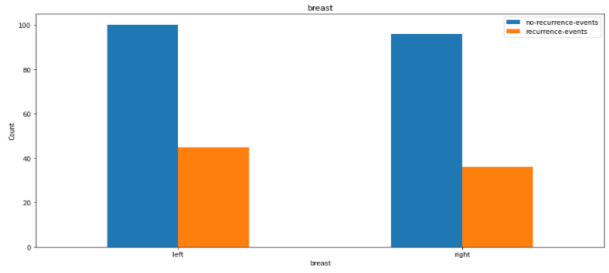

Fig. 4. The “breast” attribute

Now, it becomes clear that there are some categorical values in the dataset, and it should be replaced with Numeric values that could be processed by classification and clustering models, the replacement process was done by using the sci-kit learn Label Encoder to transfer to ordinal values. See Fig. 5.

	Class	age	menopause	tumor-size	inv-nodes	node-caps	deg-malig	breast	breast-quad	irradiat
0	0	1	2	5	0	0	2	0	1	0
1	0	2	2	3	0	0	1	1	4	0
2	0	2	2	3	0	0	1	0	1	0
3	0	4	0	2	0	0	1	1	2	0
4	0	2	2	0	0	0	1	1	3	0

Fig. 5. Label Encoder

2.4 Scaling the Dataset

Now all the dataset contains numerical values but we have to normalize our dataset because it is very important to achieve best performance in both time and accuracy. By using the Min Max normalization which scale all the data to be between [0,1]. See Fig. 6. The main idea behind normalization and standardization is the same. Variables that are measured at different scales do not contribute equally to the model fitting and model learned function and might end up creating a bias. Thus, to deal with this potential problem feature-wise normalization such as MinMax Scaling is usually used prior to model fitting. [3]

	Class	age	menopause	tumor-size	inv-nodes	node-caps	deg-malig	breast	breast-quad	irradiat
0	0.0	0.2	1.0	0.5	0.0	0.0	1.0	0.0	0.25	0.0
1	0.0	0.4	1.0	0.3	0.0	0.0	0.5	1.0	1.00	0.0
2	0.0	0.4	1.0	0.3	0.0	0.0	0.5	0.0	0.25	0.0
3	0.0	0.8	0.0	0.2	0.0	0.0	0.5	1.0	0.50	0.0
4	0.0	0.4	1.0	0.0	0.0	0.0	0.5	1.0	0.75	0.0

Fig. 6. Dataset after encoding scaling

2.5 The Correlation between the attributes

It is a commonly used method for feature selection in machine learning, and used to decide which features affect the target variable the most.[4] The correlation coefficient gives a better understanding of the linear relationship between the variables; it has a value between -1 and 1. Usually, in feature selection, the most correlated features with the target variable are important for the model training, because they provide information, no matter if this correlation is positive or negative. On other hand, we can drop the features that have a correlation close to 0. See Fig. 7.

Fig. 7. The Correlation Matrix

By looking at the above correlation matrix, it is obvious that there is not feature has a highest negative correlation; on the other hand, the "irradiate", "deg-malig", "node-caps" and "inv-nodes" features have the highest positive correlation with the target. So, these features will be considered as the most important ones for the classification model. We can notice some relationships depending on the correlation matrix, for example: "age" and "menopause" features are highly correlated, it is generally accepted that the average age at menopause is about 51 years. [5]

2.6 The Characteristics of the data

The violin plot shows the distribution of quantitative data across several levels of one (or more) categorical variables such that those distributions can be compared. [6]

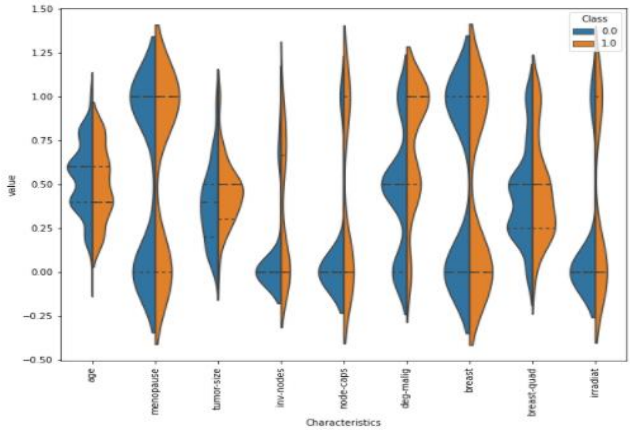

Fig. 8. Violin plot for classes' distribution

To interpret the above violin plot, we should look at the median of the label, for example, in the "tumor-size" and "deg-malig' features it's noticeable that the median of the "no-recurrence-events" and "recurrence-events" classes is separated, so it can be good for classification. However, in the other features like "breast" or "breast-quad", the median is not separated, so it doesn't give good information for classification, in other words, it isn't considered as an important feature for the model.

3 Data Preparation

After drop the "breast" attribute, the dataset will be split into two sets, training and test sets by using the train test split method from the sklearn library.

- The number of training records: 207.
- The number of testing records: 70.

The train data set, which is the largest group, will be used for training, and the test data set will be used for model evaluation.

4 Classification

Classification is considered as supervised learning, the process consists of two steps, learning step and prediction step, at the learning step the model is developed based on given training data. In the prediction step, the model is used to predict the response for

given data. All classifiers have parameters that should be set by the user called hyper parameters. [7]

The chosen models are the Gaussian Naive Bayes (Gaussian NB), Random Forest, and KNeighbors Classifier. For the evaluation part, at the end of each model, two measurements are used. The first one is "accuracy_score" from Sklearn.metrics library, it is the number of correct predictions made divided by the total number of predictions made, multiplied by 100 to turn it into a percentage. The second one is the confusion matrix, for a binary classification problem, the table has 2 rows and 2 columns. Each cell contains the number of predictions made by the classifier that fall into that cell. [8]

4.1 Gaussian Naive Bayes Classifier

In machine learning, naïve Bayes classifier is a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong independence assumptions between the features, also, it has almost no hyperparameters to tune, and so it usually generalizes well. We got accuracy to 67%. To check the cost of the model, let us see the confusion matrix Fig. 9.

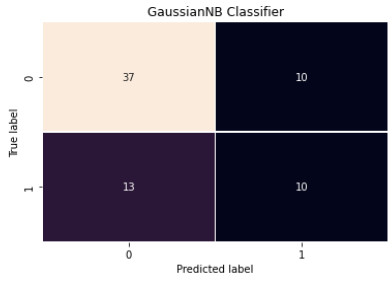

Fig. 9. Confusion matrix of the GaussianNB classifier

From the above confusion matrix, it is obvious that GaussianNB classified 23 instances in wrong way.

4.2 Random Forest Classifier

Random forest is one of the easiest and most flexible machine learning algorithms, it is a supervised learning algorithm which can be used both for classification and regression tasks, basically the forest is an ensemble of many decision trees, which merges them together to get a more accurate result. It has almost the same hyper parameters of the decision tree. [9] The most important hyper parameters of the random forest are ('n_estimators', 'max_features', 'max_depth', 'min_samples_leaf', 'min_samples_split'). First, we run the model without any tuning for the hyperparameters, and we got accuracy equal to 62%.

4.2.1 Hyper parameter Tuning

Before tuning the hyper parameters of the random forest, it's better to do cross validation, where the most common method for cross validation is K-Fold CV, which split the training set into K number of subsets, called folds. then iteratively fit the model K times, each time training the data on K-1 of the folds and evaluating on the Kth fold (called the validation data), we set K = 3 folds.

In order to find the best parameters to improve the model, and to narrow our search space, it is good to use one of RandomizedSearchCV or Gridsearchcv approaches, in our case we chose Grid search method and define a set of parameters as follows:

```
param_grid = {
    'n_estimators': [50, 100, 200, 300],
    'max_features': ['auto', 'sqrt'],
    'max_depth': [10, 20, 30, 40, 50, 60, 70],
    'min_samples_leaf': [1, 2, 4],
    'min_samples_split': [2, 5, 10]}
```

Fig. 10. Hyper parameters set

After fitting the model, we got the best parameters. See Fig. 11.

```
{'max_depth': 10,
 'max_features': 'auto',
 'min_samples_leaf': 1,
 'min_samples_split': 5,
 'n_estimators': 100}
```

Fig. 11. Random forest best Hyper parameter

4.2.2 Evaluation of Random Forest Classifier

After training the model with the best hyper parameters and running the test set on it, we got accuracy equal to 71 %, we can notice that the model accuracy got higher with 9% after tuning the hyper parameters.

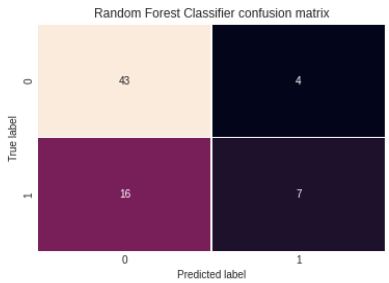

Fig. 12. Random Forest Classifier confusion matrix

From the Fig.12. The model classified 20 instances in wrong way, which less than previous model with 3 cases. Therefore, in term of accuracy and cost, the Random Forest classifier performs better than the Gaussian Naive Bayes classifier.

4.3 K-Neighbors Classifier

Nearest Neighbors (KNN) is one of the simplest algorithms used in Machine Learning for regression and classification problem, it uses data and classify new data points based on similarity measure (distance function), in other words, this classifier essentially include finding the similarity between the test pattern with all pattern in the training set. In order to find the best K to improve the mode, we will calculate the error rate for all K values between 1 and 20, and by using the plot function we get the following, see Fig.13.

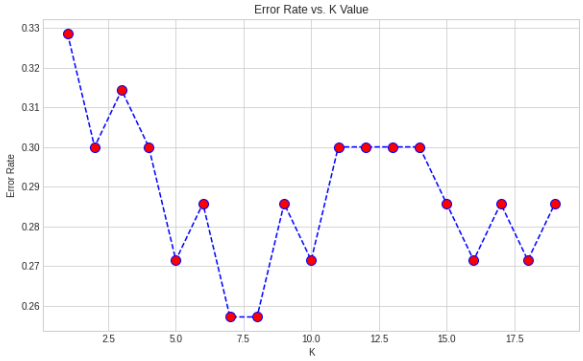

Fig. 13. The Error Rate vs. K Value

From the above figure, it is obvious that K =7 meets the lowest error rate, so it is the optimal value for the KNN model, after training the model with the best hyper parameters (k =7) and running the test set on it, we got accuracy equal to 74 %.

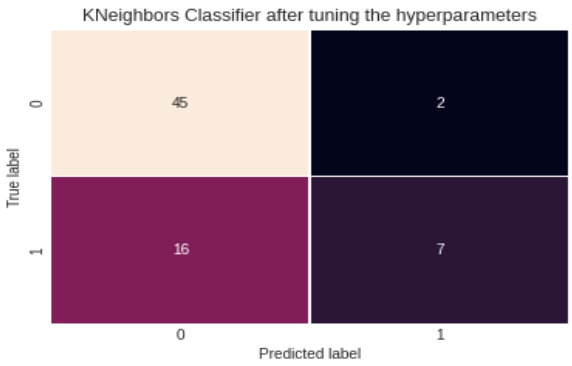

Fig. 14. KNN Classifier confusion matrix

From the Fig.14. The model classified 18 instances in wrong way, which less than previous models. Therefore, in term of accuracy and cost, the KNN classifier performs better than the previous ones.

5 Clustering

Clustering is considered as unsupervised learning, unlike the classification, here there is only the data without the labels, and the task is grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar to each other than to those in other groups clusters. [10]

5.1 KMeans algorithm

K-means clustering is an unsupervised algorithm aims to cluster\ group the data into K clusters, in other words, find k number of centroids, and then allocates every data point to the nearest cluster. K is the hyper parameter of this algorithm. First, we run the algorithm with K=2, as we already know that there are two labels in this dataset, then we will try to find the ideal K. Fig. 15. Shows the KMeans algorithm results for K =2, and two features (tumor-size, breast-quad).

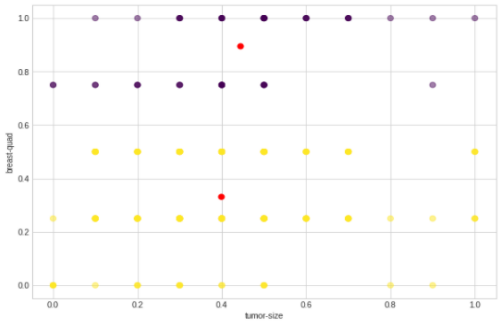

Fig. 15. K means clustering with K =2

5.1.1 Elbow Method

To compute the optimal value of k, the elbow method was used. This method fitting the model with a range of values for K, and computes the within cluster sum of squares value for each value of K. [11] By looking to the below Elbow, it shows the optimal K for the Kmeans algorithm is equal to 5.

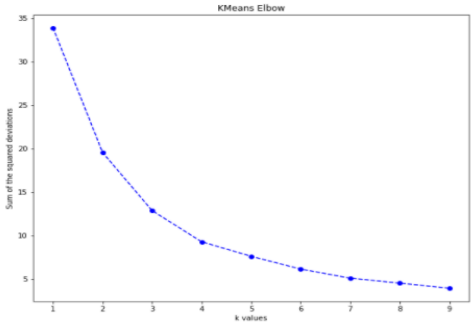

Fig. 16. Elbow Method

6 Conclusion

In this project many different algorithms and tasks are implemented and solved, namely,

- Classification problem. This study examines the ability of a set of basic machine learning methods to accurately predict the recurrence or not of breast cancer, achieving an accuracy of 74 % using the K-Neighbors Classifier. Based on this model a physician can predict recurrence of breast cancer, having to enter the following patient data: age, menopause, tumor-size, inv-nodes, node-caps, deg-malig, breast, breast-quad, irradiat.
- Clustering problem. K-means clustering algorithms is applied to study breast cancer attributes. Clustering algorithms can be combined with the decision of radiologists to improve accuracy of the cluster.

Some further approaches that can enhance the current work like in pre-processing part try to use the one hot encoding. Larger datasets can be used to increase the accuracy of the previous models. In the clustering part it could be good to try other algorithm like the Agglomerative Hierarchical Clustering, and use the principal component analysis (PCA), before fitting the model which, furthermore for the evaluation part we can use other techniques to get the optimal K.

References

[1] Coleman, M. P., Quaresma, M., Berrino, F., Lutz, J.-M., De Angelis, R., Capocaccia, R., Baili, P., Rachet, B., Gatta, G., Hakulinen, T., et al. (2008). Cancer survival in five continents: a worldwide population-based study (concord). The lancet oncology, 9(8):730–756.

[2] Mackenzie Rivero, A., Rodrıguez Rodrıguez, A., Merchan Carreno, E. J., Martınez Bejar, R. (2018). Machine Learning for the Evolutionary Analysis of Breast Cancer. Journal of Science and Research: Revista Ciencia e Investigacion, 3(CITT2017), 44-49. https://doi.org/10.26910/issn.2528 8083vol3issCITT2017.2018pp44-49.

[3] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing. MinMaxScaler.html.
[4] https://heartbeat.fritz.ai/seaborn-heatmaps-13-ways-to-customize-correlation-matrix-visualizations-f1c49c816f07.
[5] Te Velde, E. R., M. Dorland, and F. J. Broekmans. "Age at menopause as a marker of reproductive ageing." *Maturitas* 30.2 (1998): 119-125.
[6] Seaborn Page, https://seaborn.pydata.org/generated/seaborn.violinplot.html.
[7] A. Maxwell. T. Warner, F. Fang," Implementation of machine-learning classification in remote sensing: an applied review", published online: 02 Feb 2018, https://doi.org/10.1080/01431161.2018.1433343.
[8] https://en.wikipedia.org/wiki/Confusion_matrix.
[9] https://builtin.com/data-science/random-forest-algorithm.
[10] Wikipedia, https://en.wikipedia.org/wiki/Cluster_analysis.
[11] https://www.scikit-yb.org/en/latest/api/cluster/elbow.html.