Prediction models applied to lung cancer using Data Mining *

Rita Sousa[1 [0000−0002−4804−1608]], Regina Sousa[1 [0000−0002−2988−196X]], Hugo Peixoto[1 [0000−0003−3957−2121]], and José Machado[1 [0000−0003−4121−6169]]

ALGORITMI/LASI, University of Minho, Braga, Portugal
a88333@alunos.uminho.pt,
regina.sousa@algoritmi.uminho.pt,hpeixoto@di.uminho.pt,jmac@di.uminho.pt

Abstract. Lung cancer is the most common cause of cancer death in men and the second leading cause of cancer death in women worldwide. Even though early detection of cancer can aid in the complete cure of the disease, the demand for techniques to detect the occurrence of cancer nodules at an early stage is increasing. Its cure rate and prediction are primarily dependent on early disease detection and diagnosis. Knowledge discovery and data mining have numerous applications in the business and scientific domains that provide useful information in healthcare systems. Therefore, the present work aimed to compare several prediction models as well as the features to be used, with the help of Weka and RapidMiner tools. Both classification and association rules techniques were implemented. The results obtained were quite satisfactory, with emphasis on the Naive Bayes model, which obtained an accuracy of 95.03% for cross-validation 10 folds and 94.59% for percentage split 66%.

Keywords: Lung Cancer · Data Mining · Classification · Association Rules

1 Introduction

The term cancer refers to a group of diseases characterized by the development of abnormal cells anywhere in the body, which divide and grow uncontrollably [1]. Lung cancer is one of the leading cancers for both genders all over the world. It is the most common cause of cancer death and reaches 19.4% of the total [2]. Mortality and morbidity due to tobacco use is very high, about 90% of cases of lung cancer are related to exposure to tobacco smoke due to cigarettes that contains over 70 cancer-causing chemicals. This disease mostly affects people between the ages of 55 and 65 and often takes many years to develop [3]. Being a disease that is commonly misdiagnosed and a disease that impacts do many people, an early and correct diagnosis can save a lot of lives. This can be achieved with the development of predictive models of lung cancer that can help in decision

* This work has been supported by FCT—Fundação para a Ciência e Tecnologia within the R&D Units Project Scope: UIDB/00319/2020.

making software [4]. This models are developed using Data Mining approaches, that is based on the identification of anomalies, patterns, and correlations, which are difficult to find and detect with traditional statistical methods, in large data sets to predict outcomes. This work aims to compare several prediction models as well as the features to be used, with the help of Weka and Rapid Miner.

2 Related Work

V. Krishnaiah developed a paper named Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques [3], whose objective was to summarize various review and technical articles on diagnosis of lung cancer. This work compared the models are Naïve Bayes, Decision Trees (J48/C4.5), OneR and Neural Network and reached the conclusion that Naïve Bayes the most effective model at predicting patients with lung cancer disease, followed by Association Rules.

In 2015, Haofan Yang wrote a paper named Data mining in lung cancer pathology staging diagnosis: Correlation between clinical and pathology information [2]. The goal was to demonstrate the feasibility of applying the clinical information to replace the pathology report especially in diagnosing the lung cancer pathology staging. For this purpose, the Apriori algorithm was used to extract association rules and the significance of each generated rule was examined using support, confidence, and lift. The evaluation results demonstrated that the proposed *framework* can provide insight into solutions for support diagnosis of lung cancer pathological staging.

3 Materials and Methods

The dataset used was "Lung Cancer" [5] which contains data of 284 patients where 270 were diagnosed with lung cancer. To reach the main goal, a set of tools and techniques was used. Cross Industry Standard Process for Data Mining (CRISP-DM) is a structured approach to data mining and its steps were followed during the development of this work [7]. Through the use of learning or classification algorithms based on neural networks and statistics, one is able to explore a set of data, extract or help to highlight patterns and assist in the discovery of knowledge, using the Data Mining process [6]. Both Data Mining tools Weka and Rapid Miner were used in the process:
Weka is a Machine Learning workbench initially intended to aid in the application of machine learning technology to real-world data sets, specifically data sets from agricultural sector. It contains tools for data preparation, classification, regression, clustering, association rules mining and visualization [8, 9].
Rapid Miner that is a commercial tool for data analysis using machine learning that can be considered an alternative to Weka. The function of this tool is to speed up the process of creating predictive analyses and make it easier to apply them in practical business scenarios. In this tool, association rules will be performed in order to understand which symptoms manifest together[10].

4 Data Mining Implementation

Business Understanding - The main goal of this work is to develop a prediction model of lung cancer in order to assist health care professionals in making decisions to prevent diagnostic errors. As peripheral approach a set of association rules, in order to determine attribute relations was also performed.
Data Understanding - The "Lung Cancer" dataset has a cc0 license, is in the public domain and also has public visibility. This dataset has 5091 downloads and 34388 views and its source is "Online Cancer Prediction System". This dataset is composed by 16 Nominal attributes:

- Gender (M-Male, F-Female)
- Age (Integers - 21 to 87)
- Smoking, Yellow Fingers, Anxiety, Peer Pressure, Chronic Disease, Fatigue, Allergy, Wheezing, Alcohol Consuming, Coughing, Shortness of Breath, Swalloing Difficulty, Chest Pain (yes-2, no-1)
- Lung Cancer (yes, no) - label attribute.

Data Preparation

- **Weka** - Through the analysis of the attributes, it was possible to notice that there were no missing values or duplicated data. Moreover, Interquartile Range filter, showed that there were no values statistically considered outliers. Weka Attribute Selection filter, constructed a subset with the attributes age, anxiety, peer pressure, chronic disease, fatigue, allergy, wheezing, alcohol consuming, coughing, swallowing difficulty and chest pain. Since the lung cancer class had 270 patients with lung cancer and only 39 without lung cancer, it was necessary to balance the class. To do this the Smote method was used, which consists of generating synthetic data of the minority class from neighbors. Using the filter "SMOTE", it was possible to increase the number of patients that weren't diagnosed with lung cancer to 273 [11].
- **Rapid Miner** - There are values that are considered outliers in the age attribute. In order to eliminate the instances whose age corresponded to an outlier, the operator called "Delete Outlier (Distance)" was used in Rapid Miner, followed by a filter to only let through instances that are not outliers.

Modeling

- **Weka** - To achieve the prediction model a global Data Mining Model (DMM) was constructed. This is a Classification with two set of scenarios, one with all attributes and another with the attributes that resulted from applying the "Attribute Selection" filter (allergy, wheezing, alcohol consuming, coughing, swallowing difficulty and lung cancer). Data mining techniques applied were: OneR, JRip, J48, Naive Bayes and Part. In addiction, there is only one data approach which is with smote, two sample methods, cross-validation and percentage split, and one target that corresponds to the attribute lung cancer [12][13].

– **Rapid Miner** - In order to access attribute relationship, Association Rules was implemented. In order to find the most frequent item sets operator "FP Growth" was used. This operator efficiently computes all frequently occurring item sets in an ExampleSet, using the FP-tree data structure. Before adding this operator, it was necessary to transform the data that was imported as integer to binominal, since the Rapid Miner association rule operators require that the attributes of this type, by adding the "Numerical to Binominal" operator. In this operator it was necessary to create a subset consisting of the attributes whose type was wanted to be transformed. Then, the operator "Select Attributes" was added in order to reduce the number of attributes, remaining all attributes except age, gender and lung cancer. To investigate these relationships we can use the "Create Association Rules" operator. This operator uses the data from the pattern frequency matrix and looks for any patterns that occur often enough to be considered rules.

Evaluation and Discussion

– **Weka** - To evaluate the classification algorithms, performance metrics were calculated, such as accuracy, sensitivity and precision. These metrics are calculated through confusion matrices that provide the amount of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) [14]. Table 1 shows the performance metrics achieved for the algorithms. However, the OneR algorithm had the worst percentages, which can be explained by the fact that this algorithm creates rules taking only one attribute into account. The best algorithm was, by a wide margin, Naïve Bayes, although this is considered a less accurate algorithm compared to other algorithms since it assumes that the attributes are independent of each other and that there is no correlation between them [15].

Table 1. Performance metrics for the dataset consisting of all attributes

DMT	SD	Accuracy (%)	Sensitivity (%)	Precision (%)
OneR	Cross-Validation 10 folds	77,90	77,42	77,42
	Percentage Split 66%	77,84	71,20	71,20
JRip	Cross-Validation 10 folds	93,37	95,40	95,40
	Percentage Split 66%	93,51	96,59	96,59
J48	Cross-Validation 10 folds	94,29	94,49	94,49
	Percentage Split 66%	93,51	96,59	96,59
Naive Bayes	Cross-Validation 10 folds	95,21	96,25	96,25
	Percentage Split 66%	94,59	96,67	96,67
Part	Cross-Validation 10 folds	94,11	93,50	93,50
	Percentage Split 66%	94,59	94,68	94,68

The results drawn from the performance metrics for the subset, represented in Table 2, are the similar to those drawn from the dataset with all attributes, meaning that the percentages of accuracy, sensitivity and precision were high

Table 2. Performance metrics for the subset

DMT	SD	Accuracy (%)	Sensitivity (%)	Precision (%)
OneR	Cross-Validation 10 folds	77,90	77,42	77,42
	Percentage Split 66%	77,84	71,20	71,20
JRip	Cross-Validation 10 folds	94,29	95,15	95,15
	Percentage Split 66%	93,51	96,59	96,59
J48	Cross-Validation 10 folds	94,11	94,79	94,79
	Percentage Split 66%	93,51	94,56	94,59
Naive Bayes	Cross-Validation 10 folds	95,03	95,56	94,57
	Percentage Split 66%	94,59	96,67	96,67
Part	Cross-Validation 10 folds	93,74	91,93	91,93
	Percentage Split 66%	94,59	95,65	95,65

for all techniques, with the percentages for the OneR technique being the lowest and the ones for the Naïve Bayes the highest. More over, there is not much difference in results between the different sample data. Comparing the two scenarios created, it is possible to verify that there is not much difference in the values of the performance metrics. This shows that, although the "attribute selection" filter selects the attributes with the highest predictive ability, the resulting attributes are also relevant for lung cancer prediction.

– **Rapid Miner** - Since the number of association rules generated is high due to the high number of attributes, only the rules whose confidence is higher than 85% were selected. For example, the first rule of association says that if a patient's symptoms are fatigue, coughing, and wheezing, the patient in question also has shortness of breath. This rule has a support of 27.8% which means that these symptoms arise simultaneously in 27.8% of the transactions, and a confidence of 86.5% which means that in 86.5% of the transactions, patients who have fatigue, coughing and wheezing as symptoms also have shortness of breath.

5 Conclusions and Future Work

Nowadays, cancer has become devastating and is a threat to our lives. Thus, experts have introduced many useful methods to diagnose the disease at earlier stages. The high percentages of accuracy, sensitivity and precision in both sample data show that all the algorithms used are optimal to apply to the "Lung Cancer" dataset. Still, one can see that these percentages have lower values for the OneR algorithm and higher values for the Naive Bayes algorithm. The results of the association rules are ordered in increasing order of the percentage of confidence, and therefore the premise with the highest confidence is the one in the last row of the table. However, it can be seen that this is not the premise with the most support. Thus, an association could be made between rules number thirteen and fourteen because they have the most support and confidence, respectively, within the rules present in the table.

In the future, it is essential to gather more data and create new subsets to further investigate the relevance of each attribute in lung cancer.

References

1. Cancer Online, "What is lung cancer?", https://www.cancro-online.pt/cancro-do-pulmao/informacao-basica/o-que-e-o-cancro-do-pulmao/. Last accessed 5 january 2022.
2. Haofan Yang,"Data mining in lung cancer pathologic staging diagnosis: Correlation between clinical and pathology information", 2015. Last accessed 29 december 2021.
3. V. Krishnaiah, "Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques", 2013. Last accessed 2 january 2022.
4. Rita Reis, Hugo Peixoto, José Machado and António Abelha, "Machine Learning in Nutritional Follow-up Research", 2017, https://www.degruyter.com/document/doi/10.1515/comp-2017-0008/html. Last accessed 29 march 2022.
5. Mysar Ahmad Bhat, "Lung Cancer", 2021, https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer. Last accessed 18 december 2021.
6. DevMedia, "Data Mining: concepts and use cases in healthcare", https://www.devmedia.com.br/data-mining-conceitos-e-casos-de-uso-na-area-da-saude/5945. Last accessed 21 december 2021.
7. Jorge Horácio, "Data Driven Mindset – O modelo de mineração CRISP-DM", https://jorgeaudy.com/2021/01/29/data-driven-mindset-o-modelo-de-mineracao-crisp-dm/. Last accessed 22 december 2021.
8. Marcelo Damasceno, "Introduction to Data Mining using Weka", http://connepi.ifal.edu.br/ocs/anais/conteudo/anais/files/conferences/1/schedConfs//papers/258/public/258-4653-1-PB.pdf. Last accessed 30 december 2021.
9. Garner, S. R. (1995, April). Weka: The waikato environment for knowledge analysis. In Proceedings of the New Zealand computer science research students conference (Vol. 1995, pp. 57-64). Last accessed 29 march 2022.
10. iMasters, "Data Mining: Association Rules", https://imasters.com.br/back-end/data-mining-na-pratica-regras-de-associacao. Last accessed 30 december 2021.
11. Rodrigo Santana, "Dealing with Unbalanced Classes - Machine Learning", 2020, https://minerandodados.com.br/lidando-com-classes-desbalanceadas-machine-learning/. Last accessed 10 january 2022.
12. Francisca Fonceca, Hugo Peixoto, Filipe Mirande, José Machado, António Abelha, "Step Towards Prediction of Perineal Tear", 2017, https://repositorium.sdum.uminho.pt/bitstream/1822/51692/1/3.pdf. Last accessed 10 january 2022.
13. Cristina Neto, Hugo Peixoto, Vasco Abelha, António Abelha and José Machado, "Knowledge Discovery from Surgical Waiting lists", 2017, https://www.sciencedirect.com/science/article/pii/S1877050917323438. Last accessed 29 march 2022.
14. iMasters, "Machine Learning: Metrics for Classification Models", 2019, https://imasters.com.br/desenvolvimento/machine-learning-metricas-para-modelos-de-classificacao. Last accessed 11 January 2022.
15. Mariana Rodrigues, Hugo Peixoto, José Machado and António Abelha, "Understanding Stroke in Dialysis and Chronic Kidney Disease", 2017, https://www.sciencedirect.com/science/article/pii/S1877050917317052. Last accessed 29 march 2022.