engineering proceedings

MDPI

Proceeding Paper

Training of Machine Learning Models for Recurrence Prediction in Patients with Respiratory Pathologies †

Ainhoa Molinero Rodríguez [1,*], Carla Guerra Tort [1,*], Victoria Suárez Ulloa [2], José M. López Gestal [3], Javier Pereira [1] and Vanessa Aguiar Pulido [4]

[1] CITIC-Research Center of Information and Communication Technologies, University of A Coruña, 15071 A Coruña, Spain; javier.pereira@udc.es
[2] Institute for Biomedical Research of A Coruña (INIBIC)-Fundación Profesor Novoa Santos, 15006 A Coruña, Spain; victoria.suarez.ulloa@sergas.es
[3] Instituto Médico Quirúrgico San Rafael, 15009 A Coruña, Spain; jlopez@imqsanrafael.es
[4] Computational Biology, University of Miami, Miami, FL 33146, USA; vaguiarpulido@gmail.com
* Correspondence: a.molinero@udc.es (A.M.R.); c.gtort@udc.es (C.G.T.)
† Presented at the 4th XoveTIC Conference, A Coruña, Spain, 7–8 October 2021.

Abstract: Information extracted from electronic health records (EHRs) is used for predictive tasks and clinical pattern recognition. Machine learning techniques also allow the extraction of knowledge from EHR. This study is a continuation of previous work in which EHRs were exploited to make predictions about patients with respiratory diseases. In this study, we will try to predict the recurrence of patients with respiratory diseases using four different machine learning algorithms.

Keywords: electronic health record (EHR); machine learning; linear discriminant analysis; quadratic discriminant analysis; k-nearest neighbors; decision trees

Citation: Rodríguez, A.M.; Tort, C.G.; Ulloa, V.S.; Gestal, J.M.L.; Pereira, J.; Pulido, V.A. Training of Machine Learning Models for Recurrence Prediction in Patients with Respiratory Pathologies. *Eng. Proc.* **2021**, *7*, 20. https://doi.org/10.3390/engproc2021007020

Academic Editors: Joaquim de Moura, Marco A. González and Manuel G. Penedo

Published: 13 October 2021

1. Introduction

The electronic health record (EHR) is an electronic version of patient's medical history and demographic, clinical and administrative data are included in them [1,2]. The EHR was created to improve the efficiency of health systems; however, it has several applications in clinical informatics and epidemiology. Specifically, EHR have been used for patient clustering, disease prediction and pattern recognition [3].

The analysis of clinical data associated to EHRs is based in statistical and Artificial Intelligence (AI) procedures. Recently, machine learning and deep learning algorithms have been successfully used to extract informative and useful patterns from the EHRs [4].

The present study is a continuation of previous work [5] in which EHRs were exploited to make predictions about patients with respiratory diseases. In this project, we propose the use of Machine Learning to predict the recurrence of patients with respiratory diseases in less than 6, 12 or 18 months (depending on diagnosis). For this task, four machine learning algorithms were used: linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), k-nearest neighbors (kNN) and decision trees.

2. Materials and Methods

2.1. Data Set Description

Anonymous patient data were extracted from the San Rafael Hospital database. Records range from January 2000 to January 2020. The data set consisted consisted of 996 records and 40 variables. A total of 47.19% of patients suffered a relapse in less than six months, whilst 52.81% had not relapsed in that period of time.

2.2. Machine Learning Algorithms

2.2.1. Linear Discriminant Analysis

Linear discriminant analysis (LDA) is generally used to classify patterns between two classes [6]. LDA models differences among samples assigned to certain groups, in order to maximize the ratio of the between-group variance and the within-group variance.

2.2.2. Quadratic Discriminant Analysis

Quadratic discriminant analysis (QDA) is used when it is known that individual classes show distinct covariances. In this method, individual covariance matrix is estimated for every class of observations.

2.2.3. K-Nearest Neighbors

The k-nearest neighbor classifiers (k-NNCs) assumes that similar features will form a different cluster in feature space with multiple data points. The classifier takes k-nearest neighbors to find similarities between the test data and the features of a different class.

2.2.4. Decision Trees

Decision trees (DTs) are used for classification and regression. The DT predicts the value of a target variable by learning simple decision rules inferred from the data features.

3. Results and Discussion

Figure 1 shows the results obtained for the four models. The accuracy is expressed as the ratio of correctly predicted observation to the total observations; sensitivity, ratio of true positives to actual positives; and specificity, ratio of true negatives to total negatives in the data.

	LDA		QDA		KNN	Decision Trees	
	Train	*Test*	*Train*	*Test*	*Test*	*Train*	*Test*
Accuracy	63.27	59.87	64.12	61.54	57.53	93.26	60.54
Sensitivity	67.12	65.19	74.46	75.95	62.66	94.84	62.03
Specificity	58.97	53.9	50.76	45.39	51.77	91.49	58.87

Figure 1. Results obtained for the four models.

The overall accuracy for the four models is 60%; however, the accuracy value must be greater than 80% to be considered good.

The differences between sensitivity and specificity indicate that these models have a better performance predicting non-relapses than relapses. As expected, the accuracies reported by this study were lower than the ones we would expect. In this study, we used a dataset which did not have input and output parameters for a specific disease diagnostic. Clinical records from San Rafael included information about diagnosis, procedures or health system, but it did not include parameters to diagnose a respiratory disease. With aim to make better predictions, data sets need to include more useful information such as whether the patient is smoker or not, air quality or physical activity. The use of machine learning for health predictions is growing in popularity, although some challenges lie ahead.

Author Contributions: Conceptualization, A.M.R., C.G.T., J.P.; methodology, A.M.-R.; writing, A.M.R; supervision, V.S.U., J.M.L.G., V.A.P. All authors have read and agreed to the published version of the manuscript.

Funding: Centro de Investigación de Galicia CITIC and Campus Innova (agreement I+D+ 2019-20) is funded by Consellería de Educación, Universidade e Formación Profesional from Xunta de Galicia and European Union (European Regional Development Fund - FEDER Galicia 2014-2020 Program) by grant ED431G 2019/01 and Universidade da Coruña. Partially supported by the Spanish Ministry of Science (Challenges of Society 2019) PID2019-104323RB-C33.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Pham, T.; Tran, T.; Phung, D.; Venkatesh, S. Predicting healthcare trajectories from medical records: A deep learning approach. *J. Biomed. Inform.* **2017**, *69*, 218–229. [CrossRef] [PubMed]
2. Yadav, P.; Steinbach, M.; Kumar, V.; Simon, G. Mining Electronic Health Records (EHRs): A Survey. *ACM Comput. Surv.* **2018**, *50*, 85:1–85:40. [CrossRef]
3. Luo, Y.; Szolovits, P.; Dighe A.S. Using machine learning to predict laboratory test results. *Am. J. Clin. Pathol.* **2016**, *145*, 778–788. [CrossRef]
4. Shickel, B.; Tighe, P.J.; Bihorac, A.; Rashidi, P. Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. *IEEE J. Biomed. Health Inform.* **2018**, *22*, 1589–1604. [CrossRef] [PubMed]
5. Guerra Tort, C.; Aguiar Pulido, V.; Suárez Ulloa, V.; Docampo Boedo, F.; López Gestal, J.M.; Pereira Loureiro, J. Electronic Health Records Exploitation Using Artificial Intelligence Techniques. *Proceedings* **2020**, 54, 60. [CrossRef]
6. Izenman A.J. *Linear Discriminant Analysis. In: Modern Multivariate Statistical Techniques*. Springer Texts in Statistics; Springer: New York, NY, USA, 2013.