IECI 2021
The 1st International Electronic Conference on Information
01–15 DECEMBER 2021 | ONLINE
Chaired by DR. MARK BURGIN
01010 01010 01010 information
MDPI

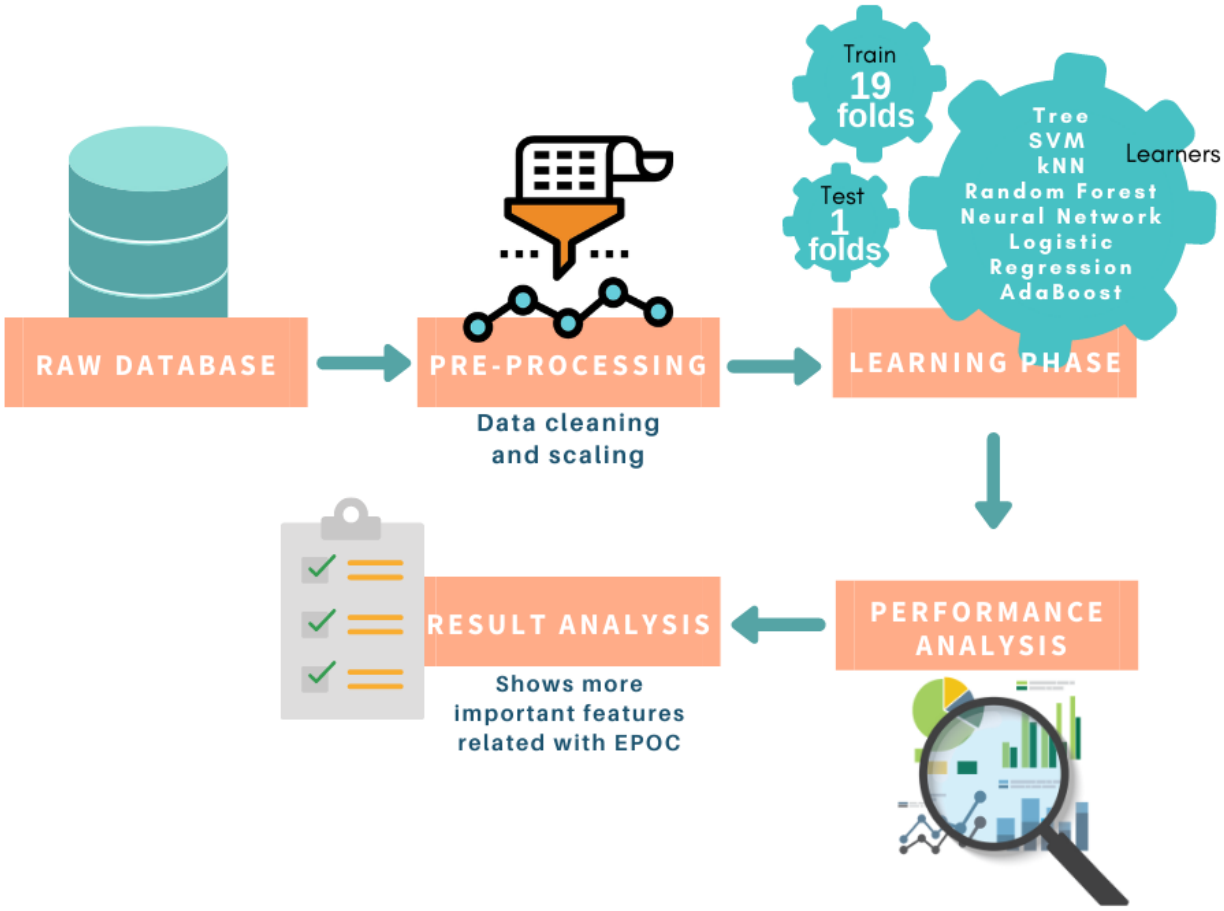
Cristián Castillo-Olea [1*], Roberto Conte-Galvan [1], Jose Juan Parcero [2], Alexandra Gómez-Siono [3] and Ornela Bardhi[4]

[1] Ensenada Center for Scientific Research and Higher Education; castillo.cristian@uabc.edu.mx, conte@cicese.mx
[2] Medica Norte Hospital ; jjparcerovaldes@gmail.com
[3] School of Engineering CETYS University, Mexicali, Baja California, Mexico 2; alexandra.siono@cetys.edu.mx
[4] Independent Researcher, Albania; alenroidhrab@gmail.com

* Correspondence: cristian.castillo2@gmail.com; Tel.: +52 (55) 7430-2237

CICESE

Prediction of Chronic Obstructive Pulmonary Disease using machine learning techniques: A Case Study from Baja California, Mexico



IECI 2021

Abstract:

Chronic Obstructive Pulmonary Disease (COPD) is a chronic inflammatory disease of the lungs that obstructs airflow from the lungs. Symptoms include difficulty breathing, coughing, mucus production, and wheezing. The study was conducted with 769 patients. A total of 48.70% were women, and 51.29% were men. The average age of the patients enrolled was 60 years. The research included 67 variables considering the medical history and biochemical data. The objective is to evaluate chronic obstructive disease. We used automatic learning techniques to assess and identify the patient's determinant variables. The following classifiers were used: Vector Support Machines (SVM), K-Nearest Neighbors (kNN), Decision Tree, Random Forest, Neural Network, AdaBoost, and Logistic Regression. The model suggests that the determining variables for COPD in treated patients are the following: TA_dist, Cholesterol level, LDL levels, Dyslipidemia, Bradycardia, Venous Insufficiency, Systolic Dysfunction, Cardiac Arrhythmia, Vasomotor Headache, Smoking, and Esophageal Achalasia. Therefore, they are considered relevant in the decision-making process for choosing treatment or prevention. The analysis of the relationship between the presence of the variables and the classifiers used to measure COPD revealed that the Logistic Regression classifier, with the variables TA_dist, Cholesterol level, LDL levels, Dyslipidemia Bradycardia, Venous Insufficiency, Systolic Dysfunction, Cardiac Arrhythmia, Vasomotor Headache, Smoking, and Esophageal Achalasia, showed an accuracy of 0.90, precision 0.87 and an F1 score of 0.89. Therefore, we can conclude that the Logistic Regression classifier gives the best results for evaluating the determining variables for COPD assessment.

Keywords: machine learning; EPOC; prediction.

Introduction

Chronic Obstructive Pulmonary Disease (COPD) is an ailment that mainly affects people with smoking habits, but also this disease has great importance in people with a hereditary history of COPD. This disease causes fast lung wear, which in turn decreases mobility and the people who have it become dependent on a daily dose of a minimum of 18 hours per day of oxygen. In 2017 COPD was the fourth cause of death as it affected about 600 million people around the world, and it is estimated that by 2030 it will become the fifth cause of disability in the world population.

IECI 2021

Materials and Methods

The database provided data about the cholesterol and LDL levels in each patients, since the values were continued, we decided to categorize them into three categories for cholesterol and two for LDL

Cholesterol Levels

Desirable: less than 200
High: 200 and 239
Very High: 240 and more

LDL Levels

Healthy: Less than 100
Not Healthy: More equal 100

IECI 2021

Materials and Methods

Machine Learning

Learners:

k-Nearest Neighbours (kNN)
Decision Tree (DT)
Support Vector Machine (SVM)
Random Forest (RF)
Multi-layer Perceptron Neural Network (MLPNN)
Logistic Regression (LR)
AdaBoost (AB)

Database description

Location: Medical Norte Hospital in Baja California, Mexico.
Sample: 769 Patients
Variables: 67

Metrics Evaluation

Parameters:
Accuracy
Precision
F1 Score
Recall
Models: 7

IECI 2021

Results

Tree classification

The first analysis after pre-processing the database was tree classification or decision tree. For this analysis, all the variables in the pre-processed database were used and the decision tree shown in its first fifth levels the following variables in Table 1.

Table 1. Decision tree variables

Target COPD	
Level 1	**Smoking habits**
Level 2	**Bradycardia**
Level 3	**Cholesterol**
Level 4	**Obesity**
Level 5	**TA_distolic**
Level 6	**Age**
Level 7	**LDL levels**

Results

Cross-validation

After the selection of the top-five variables, other features throwed by the decision tree were used iteratively and using cross validation. The combination with higher scores was selected as the dataset with important variables related to COPD. The highest scores are shown in Table 2.

Table 2. Highest scores from dataset.

Model	ACC	F1	Precision
kNN	0.799	0.729	0.749
Tree	0.865	0.851	0.869
SVM	0.835	0.826	0.859
Random Forest	0.878	0.875	0.883
Neural Network	0.873	0.873	0.869
Logistic Regression	0.900	0.899	0.870
AdaBoost	0.862	0.869	0.873

Results

Cross-validation

As seen in Table 2, the analysis revealed that the Logistic Regression algorithm got the highest scores with an accuracy of 0.900, an F1 of 0.899, and a precision of 0.870.These scores were given by using the following variables:

- TA dist
- Cholesterol levels
- LDL levels
- Dyslipidemia
- Bradycardia
- Venous Insufficiency
- Systolic Dysfunction
- Cardiac Arrhythmia
- Vasomotor Headache
- Smoking habits
- Esophageal Achalasia

Results

Cross-validation

Learner	Parameters
RF	Number of trees: 10, minimum subsets split: 5, maximum tree depth: unlimited.
kNN	Number of neighbours: 3, metric: euclidean, weight: uniform
SVM	Type: SVM Regression, C=1, Ɛ=0.1, Kernel= Radial Basis Function (RBF), exp (-auto|x-y|²), numerical tolerance: 0.001
MLPNN	Hidden layers: 100, activation: ReLu, solver: Adam, alpha: 0.0001, maximum iterations: 200, replicable training: True.
NB	Naïve Bayes
AB	Base estimator: tree, number of estimators: 50, algorithm (classification): Samme.r, loss (regression): Linear
DT	Type: binary tree, internal nodes < 5, maximum depth: 100, splitting: 95%.

Results and Discussion

There is research that studies the CT scan using automatic learning techniques, which evaluate Thoracic CT Texture, and the results are between 0.82 and 0.88 of precision [20]. A study on COPD with similar technical characteristics to our study [21] used three classifiers: Random Forest, Gradient Boosting, and Logistic Regression, with Random Forest as the best classifier, obtained an accuracy of 0.82. The article for the measurement of variables involved in the COPD, using Logistic Regression, RF and XGBoost were show results between 0.787 and 0.801 analyzed 54 variables in this study.

In our study on the diagnostic detection of COPD with 67 variables from 769 patients were analyzed, obtaining an accuracy of 0.900, precision of 0.870, and an F1 score of 0.890, which is the harmonic mean of accuracy and recovery, using the logistic regression algorithm. With this analysis, variables like diastolic blood pressure, cholesterol levels, LDL levels, Dyslipidemia, bradycardia, venous insufficiency, systolic dysfunction, cardiac arrhythmia, vasomotor headache, smoking habits, and esophageal achalasia were found closely related to COPD.

Conclusions

In this research, patients diagnosed with COPD showed how Machine Learning (ML) techniques could be used to assess respiratory biomedical variables and thus learn to identify, analyze, and suggest appropriate evidence-based treatment. The experimental data was initially sorted, evaluated, and tested. After that, it was iteratively used to train ten different models, with the Logistic Regression and Random Forest models showing the highest results of the comprehensive data sets. They revealed that the Logistic Regression algorithm got the highest scores with an accuracy (ACC) of 0.900, a F1 of 0.890, and precision 0.870. This study illustrates how ML techniques can help medicine and healthcare professionals with useful computer tools to improve and more efficiently care for lung diseases, providing more accurate diagnosis and treatment options for patients. ML-supported diagnosis can also provide physicians and healthcare personnel with accurate tools to help identify potential patients in risk, as well as to provide second-opinion diagnosis support.

Acknowledgments