

RESEARCH

Open Access



Development of glaucoma predictive model and risk factors assessment based on supervised models

Mahyar Sharifi¹, Toktam Khatibi^{1*}, Mohammad Hassan Emamian², Somayeh Sadat³, Hassan Hashemi⁴ and Akbar Fotouhi⁵

* Correspondence: toktam.khatibi@modares.ac.ir

¹School of Industrial and Systems Engineering, Tarbiat Modares University, Tehran, Iran

Full list of author information is available at the end of the article

Abstract

Objectives: To develop and to propose a machine learning model for predicting glaucoma and identifying its risk factors.

Method: Data analysis pipeline is designed for this study based on Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology. The main steps of the pipeline include data sampling, preprocessing, classification and evaluation and validation. Data sampling for providing the training dataset was performed with balanced sampling based on over-sampling and under-sampling methods. Data preprocessing steps were missing value imputation and normalization. For classification step, several machine learning models were designed for predicting glaucoma including Decision Trees (DTs), K-Nearest Neighbors (K-NN), Support Vector Machines (SVM), Random Forests (RFs), Extra Trees (ETs) and Bagging Ensemble methods. Moreover, in the classification step, a novel stacking ensemble model is designed and proposed using the superior classifiers.

Results: The data were from Shahroud Eye Cohort Study including demographic and ophthalmology data for 5190 participants aged 40-64 living in Shahroud, northeast Iran. The main variables considered in this dataset were 67 demographics, ophthalmologic, optometric, perimetry, and biometry features for 4561 people, including 4474 non-glaucoma participants and 87 glaucoma patients. Experimental results show that DTs and RFs trained based on under-sampling of the training dataset have superior performance for predicting glaucoma than the compared single classifiers and bagging ensemble methods with the average accuracy of 87.61 and 88.87, the sensitivity of 73.80 and 72.35, specificity of 87.88 and 89.10 and area under the curve (AUC) of 91.04 and 94.53, respectively. The proposed stacking ensemble has an average accuracy of 83.56, a sensitivity of 82.21, a specificity of 81.32, and an AUC of 88.54.



© The Author(s). 2021 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.