A Novel Data Mining Approach to Detect Thalassemia Patients Without Conducting HPLC Test

06 Nov.
ET06

Sadesha Balasooriya[1(*)], Kalinga Gunawardhana[1]

[1]*Department of Computing and Information Systems, Faculty of Applied Sciences, Sabaragamuwa University, Sri Lanka*

(*) Email: b.sadesha@gmail.com

The healthcare industry has a huge amount of medical data and information. But it is still not properly analyzed to discover useful information to predict future patterns. Hence, the main objective of this study was to introduce a new model to predict whether a person has a risk of having the Thalassemia disease or not.The data were collected from more than 7000 patients, who are currently participating in the HPLC test in the National Thalassemia Center at Kurunegala. The collected data were trained and tested using three different algorithms. The performance of the algorithms was evaluated using the confusion matrix. Supervised Learning Algorithms such as Decision tree (DT), Logistic Regression, Naïve Bayes were used to predict the model and Python colab online editor and Jupiter notebook were used to generate and compile the pseudo-codes. While the Decision Tree method generated 91.34% accuracy, Naive Bayes generated 68.58% accuracy for the data. In addition, Logistic Regression showed 84.39% of accuracy. Comparing these three algorithms showed that Decision Tree (DT) algorithm was the most accurate model to detect Thalassemia. According to year 2019 statistical reports of the Ministry of Health, the population of Thalassemia patients in Sri Lanka has slightly increased. Hence it can be concluded that the proposed methodology will be helpful to both doctors and patients in making proper decisions based on their health conditions.

Keywords: Decision tree (DT), logistic regression, NaÔve Bayes, confusion matrix