International Science and Technology Journal
المجلة الدولية للعلوم والتقنية
Volume 31 العدد
November 2022 نوفمبر

Analysis and Comparison of Data Mining Tools for Back and Neck Pain Prediction

Hamida Oushah[1] Samira Alshfah[2] Najwa Altheeb[2]

1.Electrical And Electronic Engineering Department / Engineering Faculty/ Sabratha University.
2.Computer Department / Faculty of Education / University of Zawia

Email: E_hamida@yahoo.com

الملخص

تُستخدم حقائب الظهر بشكل شائع بين تلاميذ المدارس لحمل الكتب والأمتعة الشخصية اليومية، من وإلى المدرسة، حيث أصبحت مشاكل العضلات والعظام المرتبطة بحقيبة الظهر مصدر قلق على أطفال المدارس، وبالتالي زيادة خطر الإصابة بآلام الظهر المزمنة في مرحلة البلوغ.

تم استخدام مجموعة بيانات واقعية للطلاب مجمعة مسبقًا حول آلام الظهر والرقبة ووزن الحقيبة المدرسية، تتكون البيانات من 11 سمة، وتظم 409 طالب من طلاب المدارس الابتدائية (204 ذكور و205 إناث) ، تتراوح أعمارهم بين (8–16) عامًا.

تم استخدام تقنيات استخراج البيانات للتحليل Weka ، Orange و Rapid miner، للتنبؤ بألم الظهر والرقبة، عن طريق خوارزمية Naïve Bayes.

كانت قيمة Recall تساوي 0.93٪ باستخدام أداة Rapid miner ، بينما Weka وOrange كانت Recall تساوي 0.75%، 0.71 ٪على التوالي. كما تشير هذه الدراسة إلى أن Precision لـ Rapid miner، Weka و Orange هي 0.74٪ ، 0.75٪ ، 0.73٪.

أظهرت هذه الدراسة أن أعلى Recall تم تحقيقها بواسطة أداة Rapid miner، بينما حققت Weka و Orange أقل إنجاز . في حين أن الدقة لهذه الأدوات متماثلة تقريبًا.

Abstract:

Backpacks are commonly used among schoolchildren, adolescents and adults for daily carrying personal belongings, from home to school, the musculoskeletal problems associated with backpack have become an increasing concern with school children, subsequently increases the risk of developing chronic back pain in adulthood.

A realistic a pre-collected data set of students on back and neck pain and school bag weight was used, with 11 attributes, consists of 409 primary school students (204 male and 205 female), age range between (8-16) years.

Data mining techniques have been used for analysis to predict back and neck pain, based on the Naïve Bayes algorithm by using three data mining tools Weka, Orange and Rapid miner.

The main observation is the Rapid miner Recall is 0.93%, while Weka and Orange Recall are 0.71%, 0.75. Also this study indicates that the Rapid miner, Weka, Orange Precision are 0.74%, 0.75%, 0.73%.

This study showed that, the highest Recall was achieved by Rapid miner tool, while the less achievement by Weka and Orange. Whereas, the Precision for these tools almost the same.

Keywords

Data mining, Naïve Bayes algorithm, Weka, Orange, Rapid miner

Introduction

Backpacks are commonly used among schoolchildren, adolescents and adults for daily carrying personal belongings, books, stationeries and laptops from home to workplaces or schools, the musculoskeletal problems associated with backpack have become an increasing concern with school children, subsequently increases the risk of developing chronic back pain in adulthood (Talbott *et al.*, 2009.; Avantika and Shalini 2013).

A realistic a pre-collected dataset of students on back and neck pain and school bag weight was used, consists of 11 attributes (Age, Sex, Class floor, Transportation, Method of carrying the bag, carrying other things, Are parents help?, Student weight in kg, the bag weight

in kg, backache or neck pain). There are 409 primary school students (204 male and 205 female), age range between (8-16) years, these data analyzed to help predict back and neck pain (Bhatla and Jyoti 2012). Research in this filed are limited. In this paper, data mining techniques have been used for analysis based on the Naïve Bayes algorithm by using three data mining tools Weka, Orange and Rapidminer (Mark Hal et al., 2008.; Zahraa, 2020).

Related work

A number of authors have written about data mining techniques in disease prediction, comparison of data mining tools, and also wrote about comparison of classification algorithms, all of which point to similar criteria for comparison and important features of data mining systems.

- Ahmed K (2017) data mining tools were compared on the basis of their classification accuracy. According to the result of three data mining tools used in this paper, it has been observed that different data mining tools give different results on the same data set using different classification algorithm. WEKA shows the best rating accuracy when compared to Rapidminer and Orange.
- Wang et al. (2008) in their research compared leading data mining software packages and some of the software quality criteria, such as reliability, modifiability, portability, ergonomics, efficiency, and comprehension. Of their research, the focus has been on the graphical user interface. Several criteria are selected as relevant, such as temperature portability of supported platforms and software architectures, and ergonomics.
- Zahraa Mohi (2020) researcher used Orange data mining tool to classify two types of selected medical data (Breast cancer and heart-disease) by applying decision tree, Naïve Bayes and K-nearest neighbor (KNN) classification algorithms. The accurately of KNN classifier was more efficient in accuracy for the both given data set while the NB classifier was the lowest efficient from the selected data classifier.

- Ramakrishnan Raman (2022), in his research used genetic algorithms (GA) to select a subset of cancer microarray data that contained a meaningful set of genes. Then, standard classifiers such as One-R, Bayesian Network, Logistic Regression, and Support Vector Machine (SVM) were developed based on these specific genes. Gene expression datasets are used to test the performance of these classifiers. According to the results of previous experiments, the combination of GA confluence and SVM is the most effective approach. In addition, the GA selection process is repeatable.
    - K.Gomathi (2016), This paper analyzed data mining techniques which can be used for predicting different types of diseases, and reviewed the research papers which mainly concentrate on predicting heart disease, Diabetes and Breast cancer, their work was focused on the early prediction of various diseases by using WEKA tool. Different data mining classification techniques (Naïve Bayes , J48) was used for the prediction of diseases and their performance was compared in order to evaluate the best classifier.
- Mariam Ali (2022), In her study, she predicted the water supply of the Euphrates using classification algorithms. And using a system that analyzes the expected supply from the upstream country, depending on several factors: rain and precipitation - temperatures - world oil price - season. The study focused on the role of modern technologies and methods and their applicability in the management and planning of water resources in general and in this study area in particular (Euphrates River).

Methodology

The methodology for this study is based on three fundamental steps: the selection of Data Mining Tools to test, the selection of Datasets to be used and the selection of classification algorithms to evaluate. The study will test techniques, and algorithms, and evaluate its classification by the accuracy metric.

Tools

The selection of the tools to test was done accordingly to the discretion of user-friendliness; all have a Graphical User Interface (GUI) without scripting. We selected to study only those that an analyst (not a programmer) is able to use as: Weka, Orange and Rapidminer.

WEKA is an open source software which is written in java and one of the most recognized data mining and machine language software (Mark Hal *et al.*, 2008).

Weka tool is an environment for executing the necessary steps in data mining, including pre-processing of the data and built a productive model. It includes algorithm and tools for clusters analysis, classification, recursion analysis, visualization and feature selection. Weka is basically a collection of machine learning

algorithms. It works fine on a multiprogramming operating system.

Orange is component-based visual programing software package for data visualization, machine learning, data mining and data analysis. Orange components are called widgets. Visual programing is implemented through an interface in which workflow are created by linking predefined or user-designed widgets. The core components of orange tool are written in C++ with wrappers in python (Rohit *et al.*, 2017).

Rapidminer is open source software which provides a good environment for data mining processes, it is written in java . It has a drag-and-drop facility which is used dataflow construction. Rapid miner able to support different file formats. As well as supports a large number of the classification and regression algorithms (Hofmann and Klinkenberg , 2013).

Classification

Data Classification is a two steps process: the training (or learning) phase and the test (or evaluation) phase where the actual class of the instance is compared with the predicted class, the classification algorithm used was Naïve Bayes.

The performance evaluation of the classifiers will be assessed by the precision and recall metric.

Naïve Bayes

Naive Bayes is a classification algorithm for multiclass classification problems.

The basis of the Naive Bayes classifier is based on Bayes' theorem. Lazy (lazy) is a learning algorithm, it can also work on unstable datasets (Aji W *et al.*, 2019), The way the algorithm works calculates the probability of each state for an element and classifies it based on the highest probability value. It can do very successful works with a little training data (Tunahan and İrem, 2021) .

Naïve Bayes is a popular model in Machine Learning applications because of its simplicity in allowing all attributes to contribute to the final decision equally.

This algorithm is used in multiple real-life scenarios such as spam filtration, Text classification, Sentiment Analysis, Recommendation System (Kaviani and Dhotre, 2017; Wibawa *et al.*, 2019).

Results and Discussion

Results

This study investigated the comparison and performance of data mining tools Weka, Orange and Rapid miner, by using the Naive Bayes algorithm to classify medical data for student dataset on musculoskeletal pain and backpack weight prediction (Bhatla and Jyoti 2012).

The classification accuracy based on precision and recall is the metric used to compare different data mining technologies, the results shown in the Table 1 and Figure.2.

Table 1: Classification accuracy of data mining tools on students on back and neck pain and school bag weight dataset.

Classification Technique	Tools	Precision	Recall
Naïve Bayes	Orange	0.73	0.75
	Weka	0.75	0.71
	Rapid miner	0.74	0.93

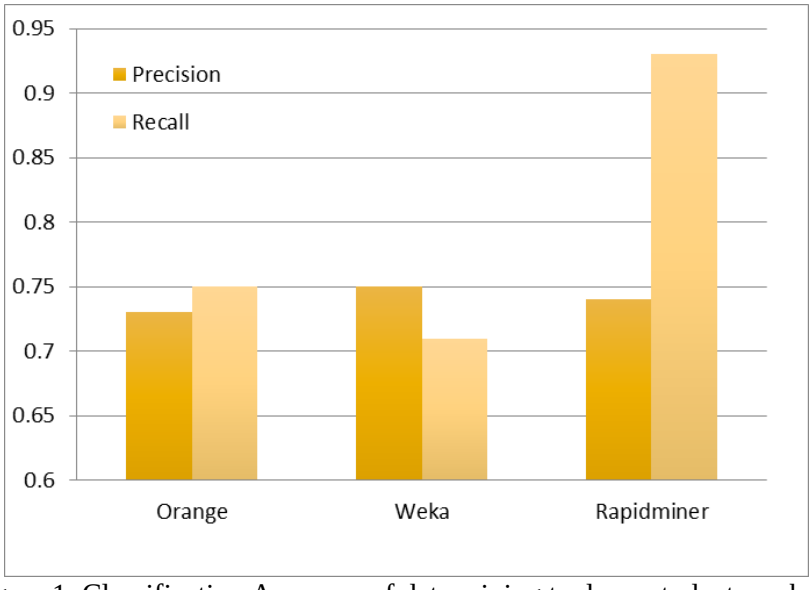

Figure.1: Classification Accuracy of data mining tools on students on back and neck pain and school bag weight dataset.

Discussion

The main observation in this study is the Rapid miner Recall is 0.93% this is almost similar to (Ahmed K, 2017), while Weka and Orange Recall are 0.71%, 0.75 which is different from (Ahmed K. 2017). Also this study indicates that the Rapid miner, Weka, Orange Precision are 0.74%, 0.75%, 0.73%, these findings are incompatible with (Ahmed K, 2017) , this is maybe due to differences of dataset nature, as well as sample size.

Similarly, (K.Gomathi *et al*., 2017) found that the accuracy by use Weka tool was 77.6 for diabetes, 79% for heart disease and 82.5% for breast cancer was almost similar.

Conclusion

In this study three different data mining techniques have been used for the prediction of back and neck pain and their performance was compared in order to evaluate the best efficiency, the highest Recall

was achieved by Rapid miner tool, while the less Recall from Weka and Orange. Whereas, the precision for these tools was similar.

References

Ahmed K (2017) Analysis of Data Mining Tools for Disease Prediction. *Journal of Pharmaceutical Sciences and Research*. 10: 1886-1888.

Aji W, Ahmad K, Della M, Risky A, Sandika P, Sulton K, Youngga N (2019) Naïve Bayes Classifier for Journal Quartile Classification. *The International Journal of Education and Science*. 2:91-99.

Avantika Rai, Shalini Agarawal (2013) Back Problems Due To Heavy Backpacks in School Children. *Journal Of Humanities And Social Science*. 6:22-26.

Bhatla N, Jyoti K (2012) An analysis of heart disease prediction using different data mining techniques. *International Journal of Engineering*. 1(8):1-4.

K.Gomathi, D.Shanmuga (2016) Multi Disease Prediction using Data Mining Techniques. *International Journal of System and Software Engineering.* 2:11-14

Kaviani P, Dhotre S (2017) Short Survey on Naive Bayes Algorithm. *International Journal of Advance Engineering and Research Development*. 11:607-611.

Mariam ali (2022) Prediction system for the Incoming quantities of the Euphrates River using machine learning Techniques. *Ministry of Higher Education & Scientific researches, Syrian Virtual University*.

MarkH, Eibe F, Geoffrey H, Bernhard P, Peter R, Ian H (2008), The WEKA Data Mining Software: An Update. *ACM SIGKDD Explorations Newsletter*. 1:10-18.

Raman R. (2022) Gene Data Analysis for Disease Detection Using Data Mining Algorithms. *Cardiometry Journal*. 25:178-181.

Rohit R, Swati A, S. Venkatesan (2017) Detailed Analysis of Data Mining Tools. *International Journal of Engineering Research & Technology*. 5: 785-789.

Talbott, N, Bhattacharya A, Davis K, Shuklab R, Levin L (2009). *School backpacks: it is more than just a weight problem*. Work. 34:481-494.

Tunahan T, İrem A (2021) Initial Seed Value Effectiveness on Performances of Data Mining Algorithms. *Journal of Science & Technology*. 9:555-567

Wang J, Hu X, Hollister K, Zhu D. (2008) A comparison and scenario analysis of leading data mining software. *Montclair State University*. 4:17–34.

Zahraa Mohi, (2020) Orange Data Mining as a tool to compare Classification Algorithms Dijlah. *Journal of Sciences and Engineering*. 3: 13-23.