sjst.scst.edu.ly
Surman Journal for Science and Technology
ISSN: Online (2790-5721) - Print (2790-5713)
مجلة صرمان للعلوم والتقنية
Vol4, No.2, Dec. 2022
Pages: 108 ~ 114

بناء شجرة القرار للتنبؤ بألم الظهر والرقبة باستخدام تقنيات التنقيب في البيانات

سميرة الشفح[1] , حميدة أوشاح[2] , إلهام أبوالشواشي[3], صديق العطاب[4]

[1] قسم الحاسوب, كلية التربية, جامعة الزاوية, ليبيا
[2] قسم الهندسة الكهربائية والإلكترونية, كلية الهندسة, جامعة صبراته, ليبيا
[3] قسم التقنية الإلكترونية, كلية التقنية الهندسية, زوارة, ليبيا
[4] كلية التقنية الهندسية صرمان, ليبيا

الملخص

يعد التنقيب في البيانات من أسرع المجالات نموا في تخصصات الحاسب الالي، ولقد جاءت شهرته وانتشاره من الحاجة المتزايدة لادوات تساعد في تحليل الكميات الهائلة من البيانات وفهمها، كما أنه يلعب دورا أساسيا في المجالات الطبية، ويمكنه حل العديد من المشاكل في سرعة التنبؤ بالامراض وصنع قرارات حكيمة.

في هذه الورقة تم استخدام التنقيب في بيانات طلبة بعض المدارس الابتدائية في مدينة صبراتة، حيث تم تحليل هذه البيانات وبناء نموذج شجرة القرار عن طريق برنامج Weka 3.8.5، للتنبؤ بمرض ألم الظهر والرقبة الناتج أثر حمل الطالب للحقيبة المدرسية.

كانت قيمة معامل الارتباط تساوي 0.0618 في المرحلة الأولى عندما قسمت البيانات 75% للتدريب و 25% للاختبار ، والخطأ المطلق النسبي =91%. بينما عندما استخدمت كل البيانات 100% لاختبار النموذج، فإن نسبة الخطأ المطلق أصبحت 73%، وقيمة معامل الارتباط = 0.45. فجودة النموذج كانت أفضل عندما استخدمت كامل سجلات البيانات في مرحلة الاختبار.

Abstract

Data mining is one of the fastest growing areas in computer disciplines, and its popularity and reach has come from the growing need for tools that help analyze and make sense of vast amounts of data. It also plays a major role in medical fields, and it can solve many problems in quick prediction of diseases and wise decision making.

In this paper, data mining was used to analyze data for some primary school students in the city of Sabratha, where this data was analyzed and then a decision tree model was built by Weka 3.8.5 software, to predict back and neck pain caused by the impact of a student carrying a school bag.

The data was divided into 75% for training and 25% for testing, and the value of the correlation coefficient in the first stage was equal to 0.0618, and the absolute relative error = 91%.

In the second stage, all data were used 100% to test the model, the absolute error rate became 73%, and the correlation coefficient value = 0.45. Model quality was better when all data records were used in the testing phase.

الكلمات المفتاحية: تنقيب البيانات، تقنيات التصنيف لتنقيب البيانات، شجرة القرار.

المقدمة

أدى التطور في توليد وجمع البيانات الى وجود مجموعات من البيانات ذات احجام هائلة في مجال الطب وكافة فروع المعرفة العلمية، حيث وجدت المؤسسات نفسها غير قادرة على ترجمة وفهم الكم الهائل من البيانات الموجودة، ولم تعد وسائل التحليل التقليدية الإحصائية قادرة على التعامل معها، فكانت تقنية التنقيب في البيانات Data mining واكتشاف المعرفة أحد الحلول الناجحة لحل هذه المشكلة.

يعتبر البعض التنقيب في البيانات مصطلحا شائعا لاكتشاف المعرفة، في حين يضع البعض التنقيب في البيانات كخطوة أساسية في عملية اكتشاف المعرفة. فلقد ظهر التنقيب في البيانات في أواخر الثمانينات والذي دخل في العديد من التطبيقات منها التطبيقات الطبية [1].

احد أساليب التصنيف المستخدمة في تنقيب البيانات هي أشجار القرار Decision Tree، والتي تقوم على بناء هيكل شجري لتمثيل القواعد المستخرجة من عملية التصنيف، حيث نالت أشجار القرار اهتماما كبيرا في كثير من المجالات واهم هذه المجالات: المجالات الطبية للتنبؤ بحالات مرضية كالتنبؤ بحدوث نوبات قلبية وتشخيص امراض السرطان بناءً على بيانات التشخيص للمرضى [2].

سنتناول في هذه الدراسة التنبؤ بألم الرقبة والظهر لطلبة المرحلة الابتدائية في مجموعة مدارس في مدينة صبراتة باستخدام تقنية شجرة القرار وخوارزمية J48، وأداة التحليل Weka 3.8.5.

المنهجية

التنقيب عن البيانات (Data Mining):

عرّف Witten وأخرون التنقيب في البيانات على أنه عملية اكتشاف أنماط جديدة من البيانات بحيث تكون هذه الأنماط ذات فائدة وجدوى، فإن استخدام تقنيات التنقيب في البيانات يوفر للمؤسسات في جميع المجالات القدرة على استكشاف والتركيز على اهم المعلومات في قواعد البيانات، كما تركز تقنيات التنقيب في البيانات كذلك على بناء التنبؤات المستقبلية واكتشاف السلوك والاتجاهات مما يسمح باتخاذ القرارات الصحيحة وفي الوقت المناسب [3].

أداة التنقيب Weka:

بدأ تطوير برنامج Weka في عام 1997، وهي أداة مفتوحة المصدر تعمل على منصة تشغيل Java، والتي تم تطويرها من قبل جامعة Waikato في نيوزلندا [4]، واستخدم في كثير من المجالات التطبيقة المختلفة خاصة للأغراض البحثية والتعليمية، وينقسم تطبيق Weka الى مجموعة من الأدوات المرتبة والخوارزميات لتحليل البيانات ونماذج التنبؤ معاً في واجهة مستخدم رسومية تسهيلا لاستخدامه، كما إنه يدعم أنواع مختلفة من تنسيقات الملفات مثل ARFF و CSV [5].

الطرق التقنية (Technical Methods):

تعتمد الطرقة التقنية على طرق تنبؤية حديثة تستخدم خوارزميات معقدة مختلفة، وهناك الكثير من هذه التقنيات مثل الشبكات العصبية (Neural Networks)، الجار القرب k-nearest neighbor، شجرة القرار (Decision Tree) [6].

شجرة القرار (Decision Tree):

شجرة القرار تقنية واسعة الانتشار تستخدم غالبا في التصنيف والتنبؤ، وهي أداة قوية لتمثيل المعرفة، وتعرف شجرة القرار بانها بنية هيكلية تستخدم لتقسيم مجموعة كبيرة من السجلات الى مجموعات صغيرة متتالية وفق تتابع معين بمخطط يشبه بناء الشجرة [7]. تستخدم أشجار القرار العديد من الخوارزميات مثل J48، M5P، ID3، والتي تقوم على التعلم من البيانات، في هذا الورقة تم استخدام خوارزمية J48 [5].

بناء النموذج:

تتكون قاعدة البيانات المستخدمة من 12 متغير (Attribute) كما موضحة في الجدول (1)، و 409 سجل (Record)، منها 204 ذكور، و 205 اناث والجدول (2) يوضح عينة من البيانات **[8].**

تمت عملية بناء النموذج باستخدام الأداة Weka الإصدار 3.8.5، وباستخدام خوارزمية J48 في شجرة القرار، وأتبع أسلوب التقسيم المئوي (percentage split) في عملية التدريب والإختبار، وهي نسبة تقسم بها البيانات ويتم تدريب واختبار النموذج عليها لاعطاء النموذج النهائي، وقد تم هنا تقسيم البيانات إلى 75% بيانات تدريب و25% بيانات اختبار، والشكل (2) يوضح شجرة القرار(Model tree) التي تم الحصول عليها، والشكل (3)، يوضح الاحصائيات لبيانات مرضى ألم الرقبة والظهر.

الجدول (1): متغيرات قاعدة البيانات

No.	Name
1	Age
2	Sex
3	Class floor
4	transportation
5	Method of carrying the bag
6	Waist belt
7	using of waist belt
8	carrying other things
9	Are parents help?
10	Student weight in kg
11	the bag weight in kg
12	Backache or neck pain

sjst.scst.edu.ly

Surman Journal for Science and Technology

ISSN: Online (2790-5721) - Print (2790-5713)

مجلة صرمان للعلوم والتقنية

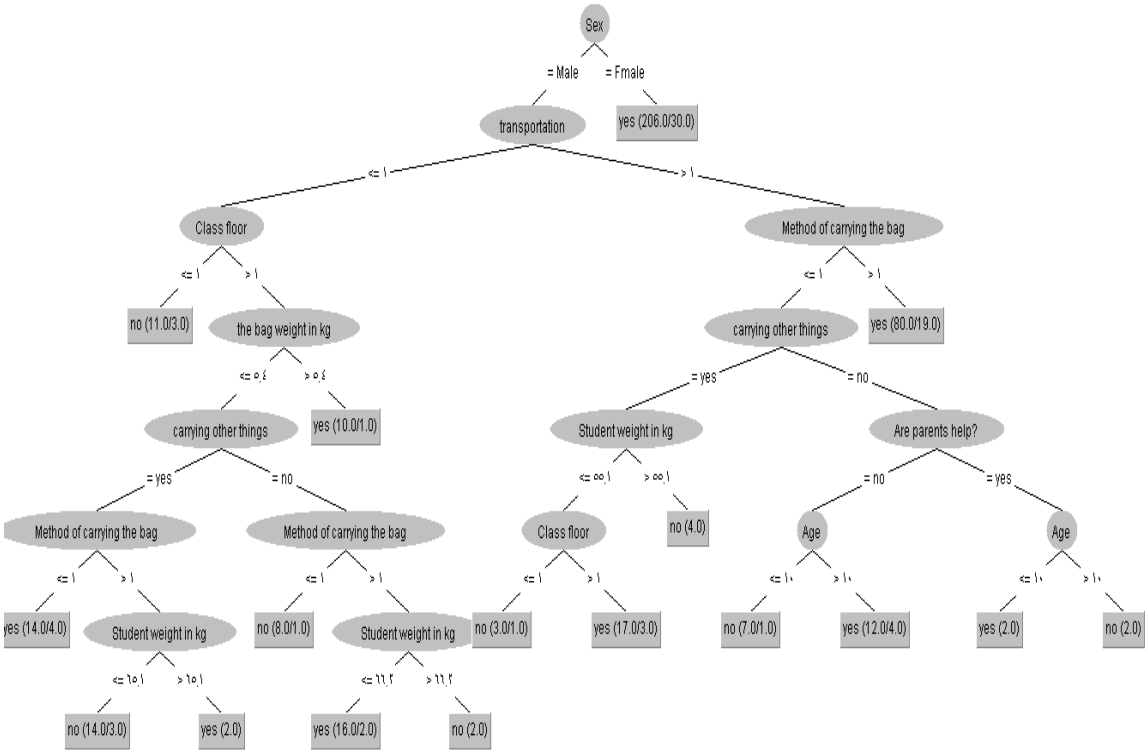
Vol4, No.2, Dec. 2022

Pages: 108 ~ 114

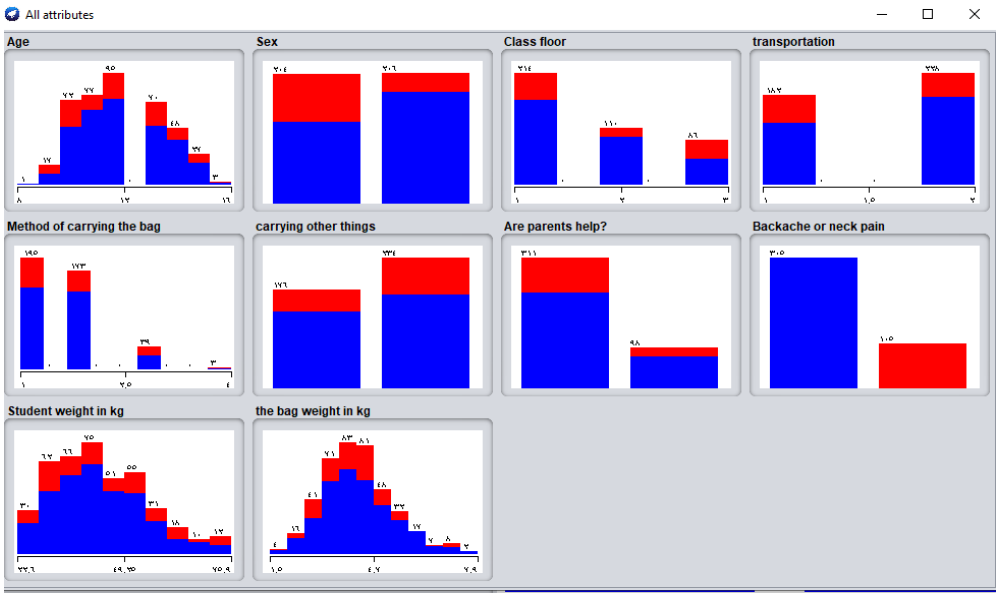
الجدول (2): بيانات الطلبة

No.	1: Age	2: Sex	3: Class floor	4: transportation	5: Method of carrying the bag	6: Waist belt	7: using of waist belt	8: carrying other things	9: Are parents help?	10: Student weight in kg	11: the bag weight in kg	12: **Backache or neck pain**
	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Nominal
1	10.0	1.0	3.0	2.0	1.0	2.0	2.0	1.0	2.0	32.8	5.1	yes
2	10.0	1.0	3.0	2.0	1.0	2.0	2.0	1.0	1.0	28.8	5.3	yes
3	10.0	1.0	3.0	1.0	1.0	2.0	2.0	1.0	2.0	31.9	4.7	yes
4	10.0	1.0	3.0	1.0	1.0	2.0	2.0	1.0	2.0	26.6	4.5	yes
5	9.0	1.0	3.0	1.0	1.0	2.0	2.0	2.0	1.0	32.4	4.2	yes
6	10.0	1.0	3.0	2.0	3.0	2.0	2.0	1.0	1.0	37.1	5.4	yes
7	10.0	1.0	3.0	1.0	1.0	2.0	2.0	1.0	2.0	32.1	4.1	yes
8	9.0	1.0	3.0	1.0	1.0	2.0	2.0	1.0	2.0	28.4	5.8	yes
9	10.0	1.0	3.0	1.0	2.0	2.0	2.0	1.0	1.0	31.1	4.3	yes
10	9.0	1.0	3.0	2.0	1.0	2.0	2.0	1.0	2.0	29.3	5.3	yes
11	10.0	1.0	3.0	1.0	1.0	2.0	2.0	1.0	1.0	28.5	4.3	yes
12	9.0	1.0	3.0	2.0	1.0	2.0	2.0	1.0	2.0	44.4	4.1	yes
13	9.0	1.0	3.0	2.0	1.0	2.0	2.0	1.0	2.0	27.7	5.1	yes
14	9.0	1.0	3.0	1.0	1.0	2.0	2.0	1.0	1.0	30.3	4.7	yes
15	8.0	1.0	3.0	2.0	1.0	2.0	2.0	1.0	1.0	27.4	4.9	yes
16	9.0	1.0	3.0	1.0	2.0	2.0	2.0	2.0	1.0	30.0	4.6	yes
17	10.0	1.0	2.0	1.0	1.0	2.0	2.0	2.0	2.0	28.7	5.6	yes
18	10.0	1.0	2.0	1.0	1.0	2.0	2.0	2.0	2.0	22.6	5.5	yes
19	11.0	1.0	2.0	2.0	2.0	2.0	2.0	1.0	1.0	26.1	3.9	yes
20	11.0	1.0	2.0	1.0	1.0	2.0	2.0	1.0	2.0	34.9	3.4	yes
21	12.0	1.0	2.0	2.0	1.0	2.0	2.0	2.0	2.0	26.4	4.5	yes
22	11.0	1.0	2.0	1.0	2.0	2.0	2.0	2.0	2.0	49.5	4.1	yes
23	11.0	1.0	2.0	2.0	1.0	2.0	2.0	1.0	2.0	30.2	4.4	yes
24	11.0	1.0	2.0	1.0	1.0	2.0	2.0	1.0	1.0	49.9	6.2	yes
25	12.0	1.0	2.0	1.0	2.0	2.0	2.0	2.0	2.0	44.4	3.5	yes
26	11.0	1.0	2.0	1.0	1.0	2.0	2.0	1.0	2.0	24.8	5.2	yes
27	11.0	1.0	2.0	2.0	1.0	2.0	2.0	1.0	2.0	36.9	3.7	yes
28	11.0	1.0	2.0	1.0	1.0	2.0	2.0	2.0	2.0	35.6	6.1	yes
29	11.0	1.0	2.0	1.0	1.0	2.0	2.0	1.0	2.0	39.3	4.1	yes
30	10.0	1.0	2.0	2.0	1.0	2.0	2.0	1.0	2.0	51.8	6.6	yes
31	11.0	1.0	2.0	1.0	2.0	2.0	2.0	1.0	2.0	67.2	5.4	yes
32	11.0	1.0	2.0	1.0	1.0	2.0	2.0	1.0	1.0	30.5	4.6	yes
33	11.0	1.0	2.0	1.0	2.0	2.0	2.0	1.0	1.0	56.1	4.9	yes
34	14.0	1.0	2.0	2.0	2.0	2.0	2.0	2.0	1.0	40.4	5.7	yes
35	11.0	1.0	2.0	2.0	2.0	2.0	2.0	2.0	1.0	31.8	4.4	yes

sjst.scst.edu.ly
Surman Journal for Science and Technology
ISSN: Online (2790-5721) - Print (2790-5713)
مجلة صرمان للعلوم والتقنية
Vol4, No.2, Dec. 2022
Pages: 108 ~ 114



الشكل (1): شجرة القرار للنموذج



الشكل (2) احصائيات تحليل بيانات مرضى الم الرقبة والظهر.

تحليل وتقييم النتائج:

النتائج التي تم الحصول عليها بتدريب واختبار النموذج كانت:

- التنبؤ بالصورة الصحيحة بنسبة 75.49% (77 حالة).
- التنبؤ بصورة خاطئة بنسبة 24.5% (25 حالة).
- قيمة معامل الارتباط تساوي 0.0618 وهو ارتباط ضعيف، حيث أنه كلما أقترب من 1 كان أفضل.
- الخطأ المطلق النسبي =91% ، تعتبر نسبة كبيرة حيث أنه كلما اقترب من 0 كان أفضل.

هذه النتائج كانت بهذه الشكل، ربما لان البيانات تم الحصول عليها من خلال إجابة الأطفال على الأسئلة فقط بدون الفحص الطبي, والأطفال عادة غير دقيقين في الإجابة, أو عند تجميع البيانات تم

سؤال الطفل مرة واحدة فقط ولم يعاد سؤاله للتأكيد، أو نتيجة لقلة التنوع في البيانات مما جعل البيانات متقاربة ويرجع ذلك لصغر حجم عينة الاختبار.

الجدول (3) يبين مصفوفة النتائج (CONFUSION MATRIX).

الجدول (3): حالات مرضى ألم الرقبة والظهر

	نعم	لا
نعم	75	3
لا	22	2

من أجل تحسين دقة النتائج ، تم ادخال كامل البيانات (410 سجل) في مرحلة الاختبار، حيث تم الحصول على نسبة خطأ مطلق 73% وهذا لا يتوافق مع دارسة أحمد [7] التي كانت نسبة الخطأ المطلق هي 22.3%.

معامل الارتباط لبيانات الدرسة = 0.45، وهذا يساوي معامل الارتباط لدراسة أحمد [7]، والذي = 0.97 وهذا الاختلاف ربما سببه بان البيانات المستخدمة في الدراسة هي بيانات دقيقة جمعت بالتحاليل الطبية، ومع هذا حدث تحسن في نسبة الخطأ المطلق ومعامل الارتباط عند ادخال البيانات كاملة وهذا لزيادة التنوع في البيانات.

الجدول (4)، يبين مصفوفة النتائج عند أخذ البيانات كاملة في مرحلة الاختبار.

الجدول (4): حالات مرض ألم الرقبة والظهر

	نعم	لا
نعم	296	9
لا	63	42

الخلاصة:

في هذه الورقة تم استخدام التنقيب في بيانات طلبة بعض مدارس التعليم الأساسي، حيث تم تحليل هذه البيانات وبناء نموذج شجرة القرار باستخدام خوارزمية J48 ، عن طريق برنامج Weka 3.8.5، للتنبؤ بمرض ألم الظهر والرقبة الناتج أثر حمل الطالب للحقيبة المدرسية. وكانت جودة النموذج أفضل عندما استخدمت كامل سجلات البيانات في مرحلة الاختبار.

المراجع:

المراجع العربية:

[1] ابوبكر محمد. تنقيب بيانات طلاب المرحمة الاساسية لمتنبؤ بدرجاتهم فى مادة الرياضيات، كلية إقرأ لدراسات الحاسوب، جامعة افريقيا العالمية، 2020.

[3] مروان ناعسة، محمد صديق. بناء شجرة قرار باستخدام خوارزمية C4.5 لدعم قرارات التسويق المباش في المصارف، مجلة بحوث جامعة حلبز العدد (23)، 2016.

[5] هيام عمر احمد. تقنيات التنقيب عن البيانات في الحقل الطبي (دراسة حالة مرض الفشل الكلوي)، كلية الدرسات العليا، جامعة النيلين، 2014.

[6] غيداء، رائف, نعمة. استخدام المصنف C4.5 في تمييز سمة الكائن دراسة مقارنة، تنمية الرافدين، العدد (28) ص 19 – 125، 2006.

[7] أحمد محمد عقاد.و بناء شجرة القرارت للتنبؤ بحجم المبيعات باستخدام تقنيات التنقيب في البيانات، كلية الاقتصاد، جامعة حلب، 2017.

المراجع الإنجليزية:

[2] Hadi , Thaer, Yazeed. Predicting Students Performance Based on their Academic Profile. Palestine Technical University Research Journal. 2:22-39, (2020).

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

[8] Waheedah Awushah. Musculoskeletal Pain and Schoolbag Use: A Cross-sectional Study Among Sabratha Student. Tripoli Medical Center (TMC), (2018).