Nile Valley University Publications
Nile Journal for Science and Engineering (NJSE)
(ISSN: ISSN :3122-5993)
Volume02, No. 01, 2026
nilevalley.edu.sd

استخدام تنقيب البيانات في التنبؤ بسرطان الثدي

دراسة حالة بمستشفى الذرة – الخرطوم

طارق عبدالكريم1 و مرشد ابراهيم2

1جامعة النيلين
2جامعة القران الكريم

المؤلف :morshedsadua@gmail.com

تاريخ الاستلام: 13 اغسطس 2025م تاريخ القبول: 17 ديسمبر 2025م

المستخلص

يقدم البحث دراسة تطبيقية عن التنقيب في البيانات واكتشاف المعرفة من البيانات الضخمة والتي غالبا ما تكون المعرفة مخفية في وسط كم هائل من البيانات "يهدف هذا البحث لاستخدام تقنيات التنقيب عن البيانات لإستكشاف المعرفة من سجلات المرضى ومعرفة اكثر الاعمار اصابة بالمرض لإجراء فحوصات وقائيةمبكرة من المرض و توفير نتائج تساهم في تقليل انتشار سرطان الثدي في الاعوام القادمة وإستخدام تقنيات تنقيب البيانات الحديثة التي تعمل علي تسهيل تحليل البيانات, تهتم الدراسة بإستخدام تقنيات تعمل علي إستخراج واكتشاف معرفة مفيدة وقابلة للاستقلال من خلال مجموعة كبيرة من البيانات . حيث نجد أن ظهور تقنيات جديدة مما ادى هذا إلي لفت الانتباه علي استخدام التقنيات في تنقيب البيانات في اكتشاف الحلول لبعض الأمراض وخاصة أمراض السرطانات حيث تضمن الدراسة في البحث في استخدام التقنيات تنقيب البيانات في سرطان الثدي باستخدام بعض الأدوات والتقنيات وتطبيق بعض خوارزميات التجميع والتصنيف باستخدام اداة التنقيب رابيدماينر وترتكز منهجية هذا البحث أولا علي تحضير البيانات التي تم الحصول عليه من مقابلة مركز الحاسوب الخاص بمستشفى الذرة ثم تطبيق تقنيات البيانات التي تم اختيارها بتسلسل معين بالرجوع إلي مجموعه من الاسباب التي يمكن تلخيصها في محورين وهما: مناسبة الطريقة لطبيعة البيانات و تلاؤمها مع أهداف البحث بالإضافة إلي كفاءة اكتشاف الأنماط مما ادى الي استخدام خوارزميات التصنيف وهي , TreeNaïve Bayes , neural networks Decisionو خوارزمية التجميع و هي K-meansلتحقيق أهداف البحث عندما تم اختيار البيانات حيث نجد أكثر الفئات العمرية عرضة للمرض هم من بين 37-46 وأكثرهم من النساء و عدد المصابين =757, حسب النوع حيث الإناث الاكثر اصابة بالمرض بعدد 6942 اصابة و الرجال الاقل بعدد 557 اصابة وان الاعوام اكثر إنتشارا للمرض حيث ان العام 2021 الاكثر انتشارا بعدد 3736 اصابة و العام الاقل 2012 بعدد 138 اصابة .

الكلمات المفتاحية : *تنقيب بيانات- البيانات الضخمة- استكشاف المعرفة - اكتشاف انماط*

The Use of Data Mining in Predicting Breast Cancer

A Case Study at Al-Durra Hospital – Khartoum

Tarig Abdelkarim[1] and Murshid Ibrahim[2]

1Neelain Univeristy

2University of Holly Guraan, Omduraman

Corresponding Author: morshedsadua@gmail.com

Received: 13th August, 2025 **Accepted: 17th Dec, 2025**

Abstract

The research presents an applied study on data mining and discovering knowledge from big data, which knowledge is often hidden in the midst of a huge amount of data.disease and provide results that contribute to reducing the spread of breast cancer in the coming years and the use of modern data mining techniques that facilitate data analysis, the study is interested in using techniques that extract and discover useful and independent knowledge through a large group of data, where we find that the emergence of new technologies. This led to drawing attention to the use of techniques in data mining in discovering solutions to some diseases, especially cancer diseases. First, I have to prepare the data obtained from the interview with the computer center of Al-Thaz Hospital Then, the application of data techniques that were selected in a specific sequence by reference to a group of reasons that can be summarized in two axes: the appropriateness of the method to the nature of the data and its compatibility with the research objectives in addition to the efficiency of pattern discovery, which led to the use of classification algorithms, namely TreeNaïve Bayes, neural networks, Decision and The aggregation algorithm, which is K-means, to achieve the objectives of the research when the data were selected, where we find that the most vulnerable age groups are among the 37-46 and most of them are women. The number of injuries was studied and distributed to the different states of Sudan. We find that the most affected states are the state of Khartoum, according to the type, where Females are the most infected with the disease, with 6942 infections, and men with the least, with 557 infections, and the years are more prevalent for the disease, as the year 2021 is the most prevalent, with 3736 infections, and the least year is 2012, with 138 infections.

Keywords: *Data mining - Big data - Knowledge exploration - Pattern discovery*

المقدمة

النموذج التلقائي **(Auto Model)** في **RapidMiner Studio**

يُعد النموذج التلقائي **(Auto Model)** امتدادًا متطورًا لبرنامج **Rapid Miner Studio**، صُمِّم لتسريع عملية بناء النماذج التحليلية والتحقق من صحتها، مع تمكين المستخدمين من تطبيقها مباشرةً في بيئة الإنتاج بدون الاعتماد على صناديق سوداء (Black Boxes)التي تُخفِي المنطق الداخلي. يُعالج هذا النموذج ثلاث فئات رئيسية من المشكلات :

1. التنبؤ بالقيم المتطرفة .(Outlier Prediction)
2. مشكلات التصنيف .(Classification)
3. مشكلات الانحدار .(Regression)

مميزات النموذج التلقائي

- تقييم البيانات :يُسهّل تحليل بياناتك وتقييمها قبل بناء النماذج .
- اقتراح النماذج :يُولّد نماذج ذات صلة تلقائيًا لحل المشكلة المحددة .
- مقارنة النتائج :يُتيح مقارنة أداء النماذج المختلفة بعد اكتمال العمليات الحسابية .
- فهم المنطق الداخلي :حتى بالنسبة للنماذج المعقدة مثل التعلم العميق (Deep Learning)، يُقدِّم تفسيرات واضحة لنتائجها، مما يُقلل من غموض "الصندوق الأسود."

واجهة Rapid Miner Studio

فيRapid Miner Studio ، يظهر النموذج التلقائي ك طريقة عرض مُخصصة (View)بجوار :

- طريقة العرض التصميمية :(Design View)لإنشاء مهام التحليل باستخدام المشغلات .(Operators)
- طريقة عرض النتائج :(Results View)لعرض مخرجات التحليل بصريًا.

تكامل التكنولوجيا والتطبيق العملي

يجمع Rapid Miner Studio بين أحدث تقنيات التنقيب عن البيانات (Data Mining) والأساليب الراسخة في واجهة واحدة سهلة الاستخدام، حيث :

- تُبنى عمليات التحليل عبر سحب وإفلات المشغلات .(Drag-and-Drop Operators)
- تُضبط المعلمات (Parameters) وتُدمج العوامل (Factors) بسهولة.

وعموما يعد Rapid Miner Studio أداة شاملة تلبي احتياجات المبتدئين والخبراء، حيث تجمع بين البساطة في الاستخدام والقدرة على التعامل مع مشكلات البيانات المعقدة(2).

مشكلة البحث

صعوبة التنبؤ والحصول على معلومات دقيقة في المستقبل وعدم الثقة لدى بعض صانعي القرار في النتائج النهائية. وإيقاع الوسائل العلمية في تحديد الأهداف والتنبؤ وإدارة العمل وكذلك ميل صانعي القرار إلي الاهتمام بالحاضر وعدم تضيع الكثير من الجهود ويضيع العديد من الفرص.

ويمكن توضيح مشكلة البحث في الاتي :

1. صعوبة استكشاف كل الاسباب المسببة للمرض بالأساليب الإحصائية التقليدية.
2. صعوبة التنبوء بإحصائيات المرض في المستقبل.
3. عدم القدرة على الاستفادة القصوى من بيانات مريض سرطان الثدي

أهداف البحث

1. إستكشاف المعرفة من سجلات المرضى باستخدام تقنيات التنقيب عن البيانات.
2. معرفة اكثر الاعمار اصابة بالمرض لإجراء فحوصات وقائيةمبكرة من المرض.
3. توفير بيانات تساعد في الحدمن انتشار سرطان الثدي.
4. التنبؤ بمدى إمكانية الاصابة بسراط الثدي بناء على بيانات المرضى باستخدام خوارزميات التصنيف .

أسئلة البحث

1-ما هي الفئات العمرية الأكثر عرضة للإصابة بسرطان الثدي؟

2- مامدى دلالة أو علاقة المهنة أو الوظيفة والحالة الاجتماعية والنوع أو الجنس والفئة العمرية بمرض السرطان؟

اهمية البحث

استخدام تقنيات تعمل علي إستخراج واكتشاف معرفة مفيدة وقابلة للاستقلال من خلال مجموعة كبيرة من البيانات .حيث يساعد في استكشاف المعرفة المخفية والنماذج غير المتوقعة ،إضافة إلى استكشاف قواعد وعلاقات جديدة موجودة في قواعد بيانات كبيرة تساعد علي معرفة اكثر الاسباب لإنتشار سرطان الثدي والعمل علي معرفة افضل طرق الوقاية وتقليل من انتشار ة المرض من ما يؤدي لتقليل نسبة الوفيات لي المرضي.

العمل علي تحليل دقيق لكمية كبيرة من البيانات المتوفرة لعدد من السنوات التي تساعد علي اتخاذ قرارات تساعد في التنبوء بمعدلات انتشار المرض في المستقبل وتوفير البيانات الازمة التي تساعد علي ارشادات ونصائح في افضل الطرق لتجنب انتشار سرطان الثدي.

وتكمن أهمية الدراسة فى أنها تتناول مرضا صار سريع الانتشار فى البلاد ولم يتم تصل الدراسات إلى القطع تماما باسباب سرعة انتشاره.

حدود البحث

1. الحدود المكانية:مستشفى الذرة بالخرطوم .
2. الحدود الزمانية: 2010الى 2021 م.

نطاق البحث

مجموعة بيانات لسرطان الثدي في الفترة من 2010-2021.

طريقة جمع البيانات

تم جمع البيانات بناء على المقابلة ، حيث تم جمعها من نظام قاعدة بيانات مستشفى "الذرة".

عينة الدراسة

تضمنت عينة الدراسة 7500 حالة مسجلة لمرضى سرطان الثدي بمستشفى الذرة.

السمات أو الخصائص

تضمنت عينة الدراسة حوالي 10 سمات أو صفات Attributesتشمل معظم البيانات الأساسية للمرضى

تجهيز البيانات

تم استكشاف البيانات Data Exploration للتأكد من سلامة وتجهيزها للمعالجة لبناء النموذج وتم استخدام الأساليب الأحصائية المتمثلة فى مقاييس النزعة المركزية كالوسط والوسيط والإنحراف المعياري والربيع الأدنى والأعلى وتم معالجة القيم المفقودة بهذه اللأساليب كما تم إزالة الضوضاء كالقيم الشاذة وغيرها وتم استخدام اسلوب التنظيف المضمن بالأدوات والبرامج المستخدة حيث أنها توفر تقنيات تنظيف تلقائية عالية الجودة.

منهجية البحث

يتبع البحث المنهج الوصفى التحليلى والتجريبي، حيث يتم جمع البيانات والمعلومات الخاصه بسجلات المراقبة وإعدادها وتصنيفها وتبويبها ومن ثم عرضها وتحليلها، ومن ثم تعتمد المنهج البنائي لبنا نموذج قادر على الاكتشاف بصوره فاعلة.

الإطار النظري

الدراسات السابقة

1.الدراسة الاولي: مدثر يونس حسن إبراهيم (2018). بعنوان: **التنبؤبمستوى الرؤية لمرض الساد باستخدام تقنيات التنقيب عن البيانات (دراسة حالة لمجمع العيون بمكة المكرمة)**. يقدم هذا البحث دراسة تطبيقية لمجال اكتشاف المعرفة باستخدام تقنيات التنقيب عن البيانات ، والهدف الرئيسي من الدراسة هو التنبؤ بمستوى الرؤية لمرضى الساد بعد العملية في مجمع عين مكة ، وكذلك معرفة العوامل التي تؤثر على الرؤية. رؤية. اشتملت الدراسة على (1452) سجلاً لمرضى أجريت لهم عملية الساد وتم الحصول عليها من المستشفى. نختار تقنية استخراج البيانات لأنه من الأفضل الاستفادة من بيانات الكمية. استخدمنا التصنيف باستخدام أشجار القرار ، وقمنا بتطبيق خوارزمية j48 على البيانات بعد المعالجة الأولية للبيانات لقاعدة البيانات ، تطبيق الخوارزميات هذا من خلال أداة weka التي تدعم المزيد من الخوارزميات وطريقة استخراج البيانات. خلصت الدراسة واستناداً إلى تحليل المريض السابق إلى أنه كان من الممكن التنبؤ بمستوى الرؤية للمرضى الجدد الذين خضعوا لعمليات الساد في وقت لاحق. من بين النتائج التي تم الحصول عليها ، تكون الرؤية بعد العملية جيدة عندما يكون المريض خاليًا من مرض السكري وارتفاع ضغط الدم ولا يزيد عمره عن 59 عامًا. وتكون الرؤية بعد العملية متوسطة عند إصابة المريض بالسكري أو ارتفاع ضغط الدم ولا يزيد عمره عن 59 عامًا. وتكون الرؤية بعد العملية سيئة عندما يكون المريض مصابًا بمرض السكر وارتفاع ضغط الدم معًا وأكثر من 59 عامًا. وخلصت التوصيات الرئيسية للدراسة إلى تطبيق الدراسة على قاعدة بيانات الساد بشكل أوسع لتشمل منطقة مريض الساد ونوع العدسة وصانع العدسة ونوع الدواء المستخدم لمعرفة تأثيره على مستوى الرؤية. (1)

2.الدراسة الثانية: شاذالي عبد الأحمد (٢٠١٧) بعنوان: **استخدام تقنيات التنقيب عن البيانات لمرضى الفشل الكلوي(دراسة حالة مستشفى احمد قاسم)**. يهدف هذا البحث إلى حل إحدى المشكلات التي يعاني منها الأطباء وهي مشكلة تشخيص أمراض الفشل الكلوي. وهناك معطيات ضخمة لا فائدة منها، لذلك جاء هذا البحث لحل هذه المشكلة بالإضافة إلى مساعدة الأطباء على اتخاذ القرار الصحيح وتقليل الإصابة بالمرض.

أجريت هذه الدراسة في مستشفى أحمد قاسم بالخرطوم على 1000 مريض منهم 590 رجلاً و 409 امرأة تتراوح أعمارهم بين 30 و 70 سنة.

تم استخدام طريقتين لاستخراج البيانات لتحليل بيانات مرضى الفشل الكلوي ، وهما تقنية التصنيف ، بما في ذلك خوارزمية j48 وتقنية التجميع ، بما في ذلك خوارزمية K-Mean لتنفيذ ذلك ، تم استخدام برامج Weak و ORANGE.

وخلصت الدراسة إلى أن الفئة العمرية والوضع الاجتماعي مرتبطان بالفشل الكلوي.

3.الدراسة الثالثة: ناهد محمد حسن أحمد (2018) بعنوان: **استخدام التنقيب عن البيانات لبناء خطط علاج لمرضى السكر**. إن وجود كميات كبيرة من البيانات عن الأمراض المزمنة أدى إلى الحاجة الملحة للاستفادة من التقنيات الحديثة لتنظيم هذه البيانات وتحويلها إلى معلومات مفيدة يمكن الاستفادة منها. في هذا البحث ، تم تقديم مشكلة تتعلق بكيفية مساعدة الأطباء على بناء خطط علاجية لتشخيص مرضى السكر باستخدام التنقيب عن البيانات. تناول البحث مرض السكري ، أنواعه المختلفة ، أسبابه ، أعراضه ، مضاعفاته ، أنواع العلاجات المتاحة ، تقنيات استكشاف البيانات الوصفية المختلفة ، التنبؤية وكيفية الاستفادة من هذه الخوارزميات في المعرفة حول مرضى السكري. تم تطوير نموذج لتشخيص الخطط العلاجية لمرضى السكر وهم المرضى الذين يتحكمون في مرض السكري وبالتالي تقل المضاعفات ويكون المرض أقل خطورة عليهم. المرضى الذين لا يسيطرون على المرض هم أكثر عرضة للمضاعفات والمرض يشكل خطرا على حياتهم. لبناء نموذج البحث ، تم استخدام مجموعة حقيقية من البيانات الطبية من المراكز الطبية، والتي تضمنت 10061 سجل طبي و 28 حقلاً. لدعم قرار الأطباء ، تم استخدام خوارزميات مختلفة للتصنيف والتجميع لبناء نموذج البحث. مر نموذج البحث بمرحلتين في المرحلة الأولى. تم تطوير نموذج تصنيف لتشخيص خطط العلاج واستخدم خوارزمية التصنيف، شجرة القرار، بايز السذاجة، اللوجيستية. بالنسبة لتحيز البيانات، تم استخدام منحنى Roc Curve لتوضيح جودة خوارزميات التصنيف. بعد عدة تجارب تم اختيار الخوارزمية اللوجستية بالنتائج: معدل الدقة 73.36، معدل الخطأ 26.64 ، 0.644 Roc، الدقة 0.696. هذه النتائج أفضل مقارنة باللوغاريتمات الأخرى (شجرة القرار، بايز ساذجة). في المرحلة الثانية ، تم استخدام نموذج تصنيف لتشخيص خطط علاج

مرض السكري واستخدمت خوارزمية العنقودية وتم استخدام متوسط K البسيط وأظهرت هذه المرحلة من النموذج دقة تصل إلى 64٪. باستخدام مرحلتين من النموذجين (التصنيف والتكتل) ، يمكن للأطباء تشخيص صحة خطط العلاج للمرضى الجدد. أوصت الدراسة باستخدام تقنية التنقيب عن البيانات في المجال الطبي لما لها من امتيازات في تقديم أفضل تشخيص للخطط العلاجية للمريض.

4.الدراسة الرابعة: هبة أحمد حسن أحمد (2018) بعنوان: **استخدام التجميع والتصنيف للتنبؤ بانتشار مرض التهاب الكبد الوبائي، دراسة حالة (ولاية الخرطوم).** هناك بيانات كبيرة ملحوظة مخزنة في قاعدة البيانات والمستودعات والتي تزداد تدريجياً. هذا الدليل لتطوير أدوات جديدة لتحليل البيانات والمعلومات / المعرفة، الاستخراج الذي يُعرف حاليًا باسم التنقيب عن البيانات الضخمة

تمثل مشكلة البحث عدم فائدة أدوات التنقيب عن البيانات في التنبؤية من الفئة العمرية المصابة وكذلك المنطقة المصابة لتسجيل البيانات والتعرف عليها

شدة مرض التهاب الكبد مقارنة بالبيئة: كان الهدف من الدراسة هو تحديد مدى انتشار التهاب الكبد الوبائي في ولاية الخرطوم ومن بين أكثر الفئات العمرية تنبؤية من خلال التنقيب عن بيانات المريض باستخدام التنقيب عن البيانات المخفي لـ قاعدة البيانات التي ستكون مفيدة للأطباء في تحديد السائد في مجالات محددة.

اعتمدت المنهجية على البيانات التي تم جمعها من وزارة الصحة- ولاية الخرطوم باستخدام أداة التنقيب عن البيانات الضعيفة بالوسائل K الخوارزمية.

أظهر أهمها انتشار وباء التهاب الكبد الوبائي في حالات ولاية الخرطوم: حيث بلغ عدد الحالات الأكثر انتشاراً 4653 حالة في الخرطوم تليها محلية الخرطوم شمال محلية أم درمان ثم محلية جبالاولية على التوالي. الأكثر فعالية للذكور من مجموعة 35-65 سنة من الإناث

5.الدراسة الخامسة: هيام عمر أحمد محمد بعنوان: **تقنيات استخراج البيانات في المجال الطبي (دراسة حالة الفشل الكلوي**. هناك العديد من الأنظمة التي تحتوي على بيانات ثمينة غامضة ، فهذه الإحصائيات من الممكن أن تعطينا الكثير من المعلومات الثمينة عند تقديمها للتحليل ولكن حجم هذه البيانات لإنشاء تحليل يدوي صعب للغاية للحصول على المعلومات المفيدة ، وبالتالي الأفضل قنوات للحصول على معلومات مفيدة من الموارد وتكنولوجيا التنقيب عن البيانات. تتناول هذه الدراسة سؤالين: ما هو المهم وأفضل خوارزمية التنقيب عن البيانات التي تستخدمها في هذا المجال (المجال الطبي) ، هل هذه الدراسة يمكن أن تساعد الإدارة في تطبيق تقنية استخراج البيانات في هذا المجال. تتمثل أهمية هذه الدراسة في كيفية استخدام الاستكشاف والتحليل بالبيانات في تكنولوجيا التعدين في المجال الطبي للحصول على المعلومات والاستنتاجات المفيدة في الدقة المرغوبة عندما يستغرق التحليل البشري أسابيع لاكتشاف معلومات مفيدة. عينة من هذه الدراسة أصبحت عامة 1120 مريض ، البيانات حول هذه العينة جمعت من قبل الباحث.

تهدف هذه الدراسة إلى التنبؤ بنوع الفشل الكلوي. حيث يتم بناء قاعدة البيانات من تاريخ المريض والمعلومات الطبية للمريض بعد تحليل البيانات الخافتة حول برنامج WEKA الذي يتنبأ به بواسطة الخوارزمية C4.5 ، تنص التنبؤات (الفشل الكلوي المزمن ، الفشل الكلوي الحاد) هذه الخوارزمية هي الأفضل للتنبؤ بنوع الكلى خزي. ووجدت الدراسة عوامل تأثير نوع الفشل الكلوي تشمل (المسببات ، الحالة ، فرط التوتر).

اكتمل بناء النموذج بشجرة القرار ، وأخيراً بلغت دقة النموذج 74٪ مع معدل خطأ 0.35(5).

مقارنة الدراسات السابقة

1. مقارنة دراستنا بالدراسة الأولى لأن دراستنا ركزت على Crispmethodology وأداة RapidMiner واستخدام تقنيات استخراج البيانات (التجميع والتصنيف) واستخدام الخوارزميات ومقارنتها مع بعضها البعض. الخوارزميات المستخدمة في دراستنا هي الشبكات العصبية وأشجار القرار. بينما ركزت الدراسة الأولى على منهجية الوصف التحليلي لوصف وتحليل البيانات باستخدام أداة Weka ، وباستخدام تقنيات الاستكشاف ، وكانت مختلفة في بعض المشكلات والأهداف والتوصيات بين الدراستين.

2.مقارنة دراستنا بالدراسة الثانية لأن دراستنا ركزت على منهجية Crisp وأداة RapidMiner واستخدام تقنيات استخراج البيانات (التجميع والتصنيف) واستخدام الخوارزميات ومقارنتها مع بعضها البعض. الخوارزميات المستخدمة في دراستنا هي الشبكات العصبية وأشجار القرار. بينما ركزت الدراسة الثانية على المنهج الوصفي التجريبي ، فقد اعتمدت على إعداد وتصميم وتطبيق تجربة عملية لاستخراج البيانات. تصف هذه التجربة وتناقشها ، وتم استخدام خوارزمية التصنيف باستخدام أداة Wicca.

3. مقارنه دراستنا بالدراسة الثالثة :ركزت هذه الدراسة على استخدام النهج التحليلي الوصفي باستخدام أداة التنقيب عن البيانات WEKA وتحليل بيانات المريض. والتوصية بتطبيق خوارزمية قواعد التباعد مطابقة مع دراستنا

4.مقارنة دراستنا بالدراسة الرابعة لأن دراستنا ركزت على منهجية Crisp وأداة RapidMiner واستخدام تقنيات استخراج البيانات (التجميع والتصنيف) واستخدام الخوارزميات ومقارنتها مع بعضها البعض. الخوارزميات المستخدمة في دراستنا هي الشبكات العصبية وأشجار القرار. بينما ركزت الدراسة الرابعة على منهجية الوصف التحليلي للبيانات باستخدام أداة Wicca ، وباستخدام تقنيات التنقيب عن البيانات في هذه الدراسة ، تم استخدام التجميع والتصنيف والتنبؤ باستخدام خوارزميات شجرة القرار. وكانت متشابهة في بعض المشاكل والأهداف والتوصيات بين الدراستين

5. مقارنة دراستنا بالدراسة الخامسة لأن دراستنا ركزت على منهجية كريسب ؛ ودراسة أخرى باستخدام التنبؤ بالخوارزمية C4.5 استخدمت دراستنا أداة RapidMiner واستخدام تقنيات استخراج البيانات (التجميع والتصنيف) واستخدام الخوارزميات ومقارنتها مع بعضها البعض حيث بنيت الدراسة من تاريخ المريض والطب معلومات للمريض بعد تحليل البيانات الخافتة حول برنامج WEKA ؛ اكتمل بناء النموذج بشجرة القرار ، وأخيراً بلغت دقة النموذج 74٪ مع معدل خطأ 0.35

الجدول التالي جدول 1-1 يوضح مقارنة بين هذه الدراسات والدراسة الحالية

جدول 1-1 يوضح مقارنة بين هذه الدراسات والدراسة الحالية

الدراسة	المنهجية	الأدوات	التقنيات	الخوارزميات	أوجه الشبه	أوجه الاختلاف	المشكلات الأهداف التوصيات
الدراسة الحالية	منهجية CRISP	RapidMiner	التجميع، التصنيف	الشبكات العصبية، أشجار القرار	-	-	تحليل بيانات المرضى، تحسين دقة النماذج، توصيات بتطبيق خوارزميات مُحسنة.
الدراسة (1)	منهجية وصفية تحليلية	Weka	استكشاف البيانات	غير محددة	التركيز على تحليل البيانات.	(Weka vs. RapidMiner) ` اختلاف التقنيات (استكشاف vs.تجميع/تصنيف) اختلاف الأهداف والتوصيات.	تحليل البيانات العامة، توصيات عامة.
الدراسة (2)	منهجية وصفية تجريبية	Weka	تصميم تجربة عملية	التصنيف (أداة غير محددة)	استخدام تقنيات التصنيف.	-اختلاف المنهجية) تجريبية .vs .(CRISPاختلاف التقنيات)تجربة عملية .vs مقارنة خوارزميات.(	تطبيق عملي لاستخراج البيانات، مناقشة نتائج التجربة.
الدراسة (3)	منهجية تحليلية وصفية	WEKA	تحليل بيانات المرضى	قواعد الترابط	التركيز على بيانات المرضى.	-اختلاف الخوارزميات (قواعد الترابط .vs شبكاتعصبية/أشجار قرار .(توصيات متشابهة في استخدام خوارزميات محددة.	تحليل بيانات المرضى، توصية بتطبيق قواعد الترابط.
الدراسة (4)	منهجية وصفية تحليلية	Weka	التجميع، التصنيف، التنبؤ	أشجار القرار	استخدام تقنيات التجميع والتصنيف.	-اختلاف الأدوات .vs Weka) .(RapidMinerتشابه في بعض الأهداف (تحسين النماذج).	تحليل بيانات عام، أهداف متشابهة جزئيًا.
الدراسة (5)	منهجية CRISP	WEKA	التصنيف (بيانات المرضى)	أشجار (C4.5 (قرار	استخدام منهجية CRISP وتحليل بيانات المرضى.	-اختلاف الأدوات، اختلاف الخوارزميات .vs C4.5) شبكات عصبية - .(دقة النموذج (74% في الدراسة نتائج الدراسة الحالية غير مذكورة)	بناء نموذج دقة 74%، توصيات بتحسين الأداء.

ملاحظات عامة على مقارنة الدراسة مع الدراسات السابقة

1. الدراسة الحالية تتميز باستخدام منهجية **CRISP**المتكاملة مع أداة **RapidMiner**، مما يوفر إطارًا منظمًا لاستخراج البيانات.
2. معظم الدراسات السابقة تعتمد على **Weka**، بينما تستخدم الدراسة الحالية **RapidMiner**، مما يشير إلى توجه مختلف في اختيار الأدوات.
3. تشابهت الدراسة الحالية مع الدراسة (5) في استخدام منهجية **CRISP**، لكنهما اختلفتا في الخوارزميات والأدوات.
4. الدراسات (1)، (2)، (4) اختلفت في المنهجية والتقنيات، مما يبرز تميز الدراسة الحالية في المقارنة بين خوارزميات متعددة.

مفاهيم تنقيب البيانات

مقدمة عن تنقيب البيانات

- تعريفه:تقنية تعتمد على خوارزميات رياضية لاستخراج المعرفة من كميات كبيرة من البيانات، مستمدة من علوم مثل الإحصاء، الذكاء الاصطناعي، وتعلم الآلة.
- تاريخه:ظهر في الثمانينات كحل لتحليل البيانات الضخمة وتحويلها إلى معلومات قابلة للاستخدام.
- التحديات:التوسع السريع في الخوارزميات والبرمجيات جعل تتبع التقنيات المتاحة أمرًا معقدًا.

المصطلحات الأساسية

- البيانات:**(Data)** حقائق وأرقام خام قابلة للمعالجة.
- المعلومات:**(Information)** علاقات ونماذج مُشتقة من البيانات.
- المعرفة:**(Knowledge)** تحويل المعلومات إلى أنماط تاريخية أو تنبؤات مستقبلية.
- مستودعات البيانات:مصممة لتحليل البيانات الزمنية واتخاذ القرارات، وتخزين بيانات من مصادر متنوعة.

استخدامات تنقيب البيانات

- تحليل العوامل الداخلية (مثل الأسعار) والخارجية (مثل المنافسة) في الأعمال التجارية.
- مثال تطبيقي:متجر كبير يستخدم التنقيب لربط مبيعات الحليب والخبز، مما يساعد في ترتيب المنتجات لزيادة الأرباح.

أمثلة عملية

- في المطاعم: تحليل طلبات الزبائن لتحديد الوجبات الأكثر شيوعًا.
- في متاجر السفر: اكتشاف أن مشتري الأكياس النوم غالبًا ما يشترون حقائب الظهر.

مراحل اكتشاف المعرفة

1. اختيار البيانات :تحديد مصادر البيانات المناسبة.
2. تهيئة البيانات :تنظيفها ومعالجة القيم المفقودة أو المكررة.
3. تحويل البيانات: تهيئتها لتتناسب مع الخوارزميات.
4. التنقيب :تطبيق تقنيات ذكية لاستخراج الأنماط.
5. تقييم الأنماط: قياس فاعليتها بناءً على أهداف المشكلة.
6. تمثيل المعرفة: عرض النتائج بطرق مرئية لتسهيل الفهم.
7. **مراحل عملية التنقيب**
8. فهم طبيعة العمل: تحديد الأهداف بوضوح (مثل زيادة الأرباح).
9. فهم البيانات:تحليل طبيعة البيانات لاختيار الخوارزميات المناسبة.
10. استخدام مستودعات البيانات :يُفضل استخدامها إذا كانت متاحة، لكنها ليست ضرورية دائمًا.
11. القيود والتحديات التي قد تؤثر على دقة التنبؤات فى تنقيب البيانات(4).

جودة البيانات **(Data Quality)**

- البيانات الناقصة: **(Missing Values)**

قد تؤدي الثغرات في البيانات إلى نتائج غير دقيقة، خاصة إذا تعاملت الخوارزميات معها بشكلٍ غير صحيح (مثل إهمال الصفوف أو ملء القيم بشكل عشوائي).

- الضوضاء: **(Noise)**

الأخطاء العشوائية أو التباين غير الضروري في البيانات (مثل أخطاء القياس) قد تُضعف قدرة النموذج على استخراج الأنماط الحقيقية.

- القيم الشاذة: **(Outliers)**

البيانات المتطرفة قد تُحرف النتائج، خاصة في نماذج مثل الانحدار الخطي أوK-means.

اختيار الخوارزمية **(Algorithm Selection)**

- التناسب مع طبيعة البيانات

بعض الخوارزميات (مثل الشبكات العصبية) تحتاج إلى كميات هائلة من البيانات، بينما أخريمثلDecision Treesتعمل بشكل أفضل مع بيانات صغيرة.

- التحيز والتباين: **(Bias-Variance Tradeoff)**

النماذج البسيطة قد تعاني من تحيز عالٍ (Underfitting)، بينما النماذج المعقدة قد تعاني من تباين عالٍ .(Overfitting)

توازن البيانات **(Class Imbalance)**

- في مسائل التصنيف، إذا كانت إحدى الفئات ممثلة بشكل ضعيف (مثل احتيال مالي بنسبة 1%)، قد تفشل النماذج في اكتشاف الفئة النادرة .
    - الحلول :استخدام تقنيات مثل SMOTE أو تعديل أوزان الفئات

التغييرات الزمنية **(Concept Drift)**

- تغير توزيع البيانات مع الوقت (مثل تغير سلوك العملاء) قد يجعل النموذج المدرب على بيانات قديمة غير دقيق .
    - التحدي :تتبع التغييرات وإعادة تدريب النموذج باستمرار.

القيود الحسابية **(Computational Constraints)**

- حجم البيانات :قد تفشل الخوارزميات التقليدية في التعامل مع مجموعات البيانات الضخمة (Big Data) دون تحسينات مثل التوازي .(Parallel Processing)
- الوقت :بعض النماذج) مثل SVM مع مجموعات البيانات الكبيرة (تتطلب وقتًا طويلًا للتدريب.

تفسير النتائج **(Interpretability)**

- النماذج المعقدة (مثل الشبكات العصبية) توفر دقة عالية لكنها تفتقر إلى الشفافية ("صندوق أسود") .
- قد تكون النماذج البسيطة (مثل الانحدار اللوجستي) أقل دقة لكنها أسهل في التفسير، خاصة في المجالات الحرجة (مثل الطب).

الخصوصية والأمان **(Privacy & Ethics)**

- قيود مثل اللائحة العامة لحماية البيانات (GDPR) قد تحد من استخدام بعض البيانات الحساسة .
- التحيز في البيانات (Bias) قد يؤدي إلى نتائج غير عادلة (مثل تحيز جندري أو عنصري).

التكامل مع الأنظمة الحالية **(Integration)**

- صعوبة دمج نماذج التنقيب مع الأنظمة القديمة .(Legacy Systems)
- عدم توافق تنسيقات البيانات بين المصادر المختلفة) مثل قواعد البيانات العلائقية وملفات.(JSON

التحديات البشرية **(Human Factors)**

- نقص الخبرة :قد تُستخدم الخوارزميات بشكلٍ خاطئ بسبب سوء الفهم .
- التعاون بين الفرق :قد تختلف أولويات فرق البيانات (Data Scientists) وفرق الأعمال.(Business Analysts)

تقييم النموذج **(Model Evaluation)**

- اختيار المقاييس الخاطئة) مثل الاعتماد على الدقة Accuracy في البيانات غير المتوازنة .(
- الاعتماد على بيانات الاختبار (Test Data) غير الممثلة لتوزيع البيانات الحقيقية.

الحلول المقترحة

- تحسين جودة البيانات :استخدام تقنيات التنظيف (Data Cleaning) وملء الثغرات بذكاء.
- التحقق المتقاطع : (Cross-Validation) لتقليل التأثر بالتوزيع العشوائي للبيانات .
- ال : regularization_للحد من .Overfitting
- التعلم التعزيزي : (Ensemble Learning) مثل Random Forests أو Gradient Boosting لتحسين الدقة .
- المراقبة المستمرة :لاكتشاف Concept Drift وإعادة تدريب النموذج. (3)

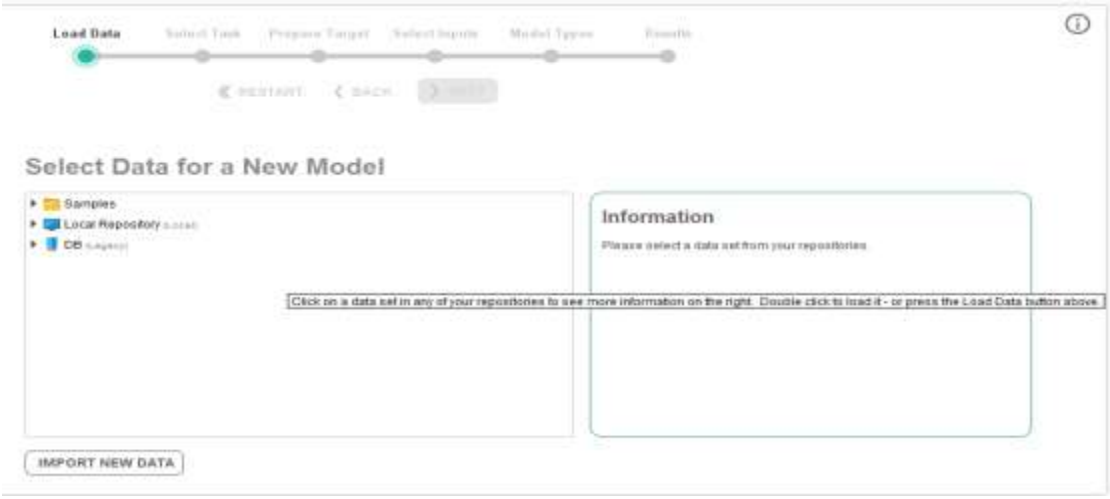
تطبيق خوازميةالتصنيف وهي شجرة القرار(Tree Decision)

1- مجموعة البيانات وخصائص الدراسة وانواعها

الجدول رقم (1) يوضح خصائص البيانات

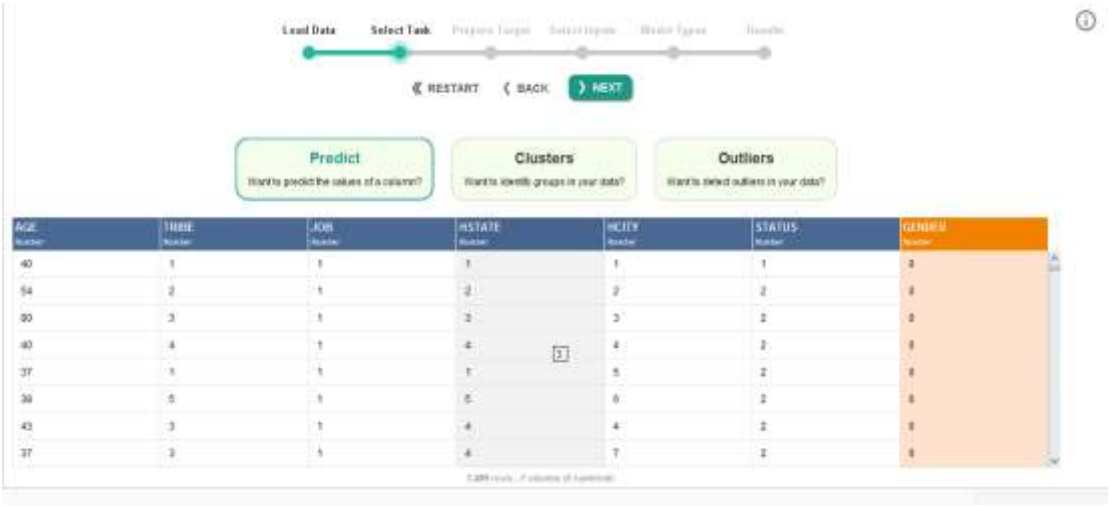
Attributes	Field name
NUMERIC	AGE
STRING	GUNDER
STRING	TRIBE
STRING	JOB
STRING	HSTATE
STRING	HCITY
STRING	STATUS
DATE	NEWCASE_DATE

1. الشكل يوضح المرحلة الاولى من Auto Model حيث يتم إختيار البيانات



الشكل رقم (1)المرحلة الولى (Auto Model) .

الشكل 2 يوضح المرحلة الثانية اختيار الفئة و تم اختيار التنبؤ.

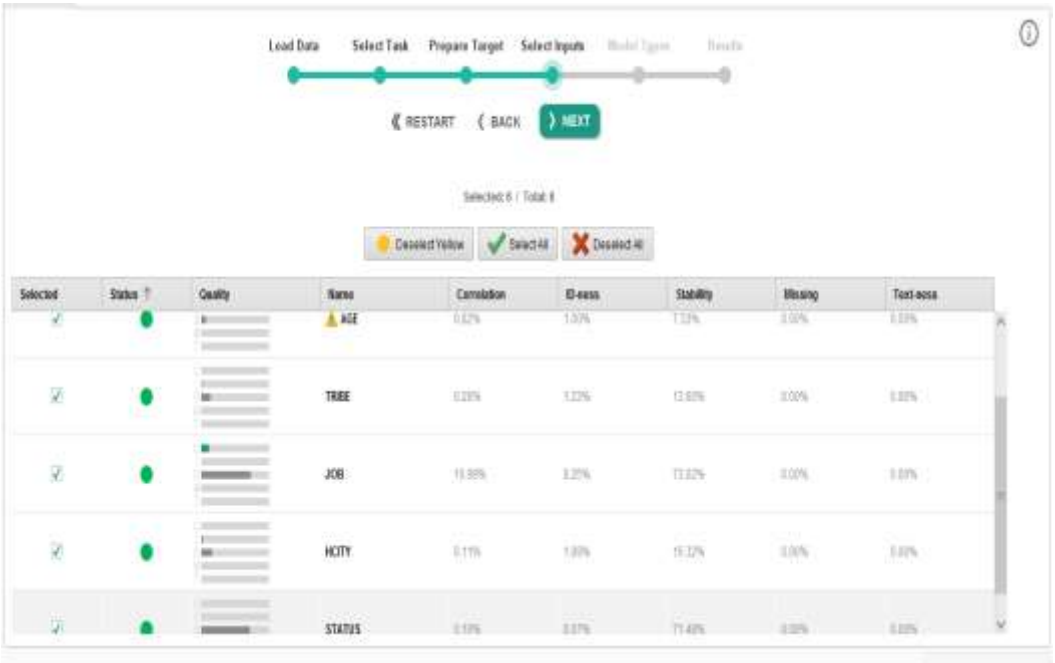


الشكل رقم (2) اختيار الفئة (Auto Model) .

الشكل 3 يوضح المرحلة الثالثة تم تقسيم البيانات حسب عدد الكلاس حيث حدد عدد الاناث كان قرابة 7000 و عدد الزكور 500

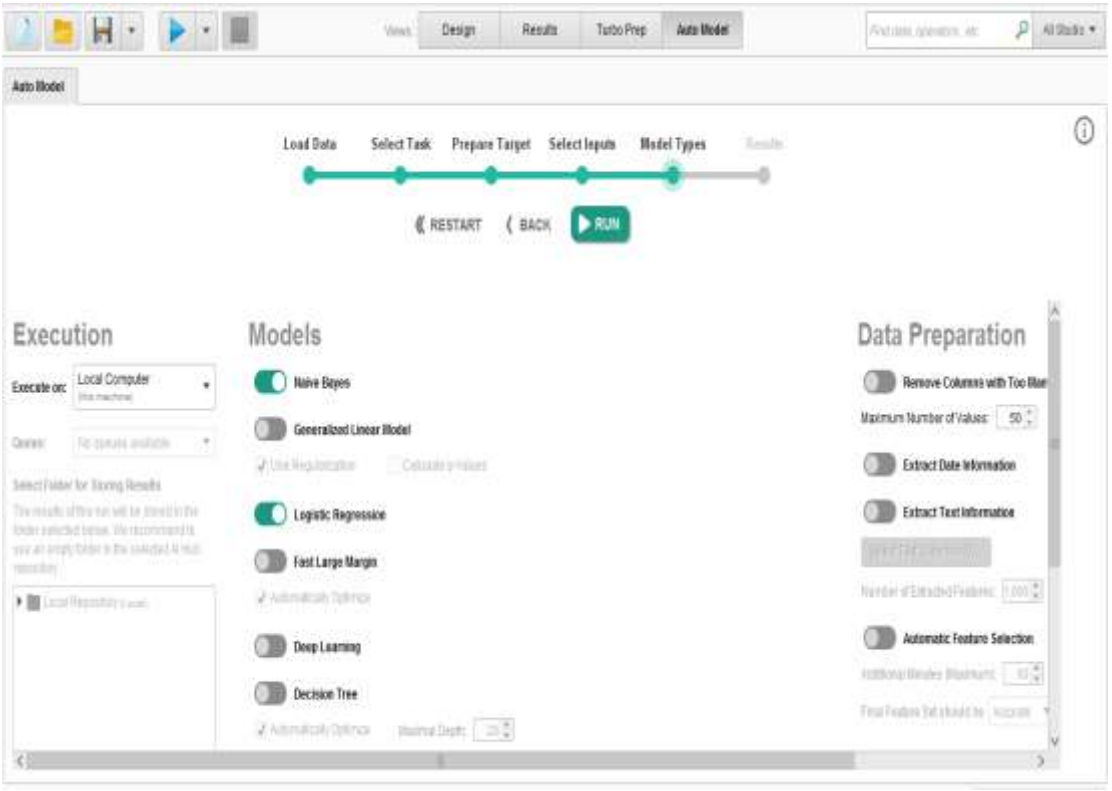
الشكل رقم (3) تقسيم البيانات (Auto Model) .

الشكل 4 يوضح المرحلة الرابعة و يتم إختيار الخصائي يتم تنفيذها



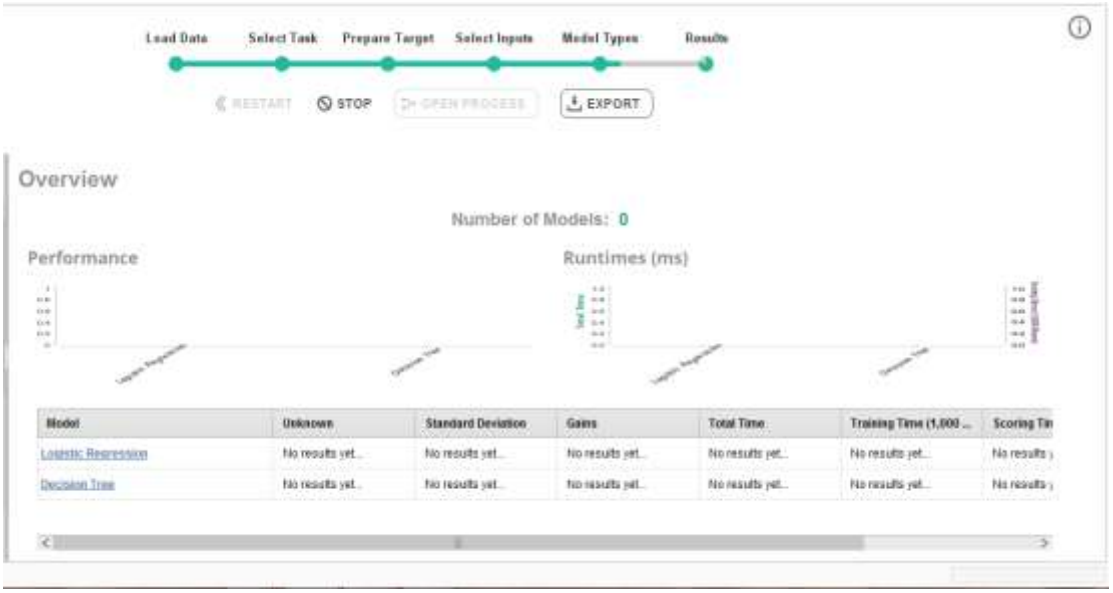
الشكل رقم (4) يوضح اخيار الخصائص(Auto Model) .

الشكل 5 يوضح المرحلة الخامسة اختيار الخوارزمية المراد تطبيقها.



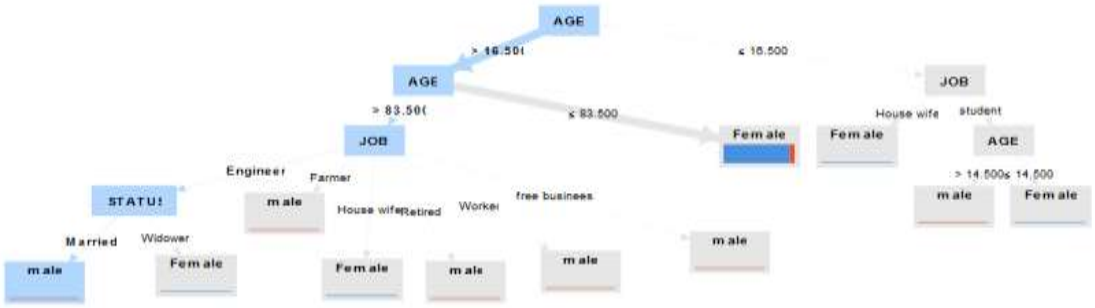
الشكل رقم (5) اختيار الخوارزمية(Auto Model).

الشكل 6 يوضح المرحلة الاخيرة وهي عرض النتائج



الشكل رقم (6) عرض النتائج(Auto Model).

شجرة القرار :يوضح الشكل 7 شكل شجرة القرار والعلاقات بين الحقول:



الشكل رقم (7) يوضح شجرة القرار المصدر برنامج(Rapid Miner)

يوضح الشكل 8 نسب ارتباطات شجرة القرار بين الحقول ونسبة توزيعها على كل فئة

Tree

```
AGE > 16.500
|   AGE > 83.500
|   |   JOB = Engineer
|   |   |   STATUS = Married: male {Female=0, male=5}
|   |   |   STATUS = Widower: Female {Female=3, male=0}
|   |   JOB = Farmer: male {Female=0, male=8}
|   |   JOB = House wife: Female {Female=64, male=0}
|   |   JOB = Retired: male {Female=0, male=7}
|   |   JOB = Worker: male {Female=0, male=2}
|   |   JOB = free business: male {Female=0, male=3}
|   AGE ≤ 83.500: Female {Female=6866, male=526}
AGE ≤ 16.500
|   JOB = House wife: Female {Female=6, male=0}
|   JOB = student
|   |   AGE > 14.500: male {Female=0, male=6}
|   |   AGE ≤ 14.500: Female {Female=3, male=0}
```

الشكل رقم (8) يوضح وصف قواعد شجرة القرار المصدر برنامج (Rapid Miner).

العدد النسبي للأمثلة المصنفة بشكل صحيح أو بعبارة أخرى النسبة المئوية للتنبؤات الصحيحةو دقة الخوارزمية هي.**92.53%**.


Table View Plot View

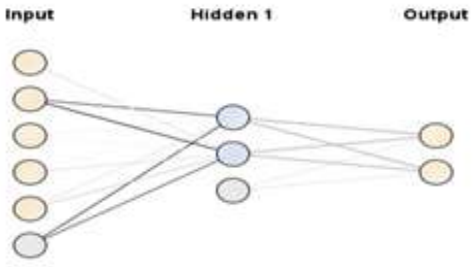
accuracy: 92.53%

	true Female	true male	class precision
pred. Female	2078	163	92.73%
pred. male	5	4	44.44%
class recall	99.76%	2.40%	

الشكل رقم (9) يوضح النسبة المئوية للتنبؤات الصحيحة.المصدر برنامج (Rapid Miner).

الشبكة العصبية :

الشبكة العصبية تتكون من مجموعة من وحدات الحوسبة الأولية المعروفة باسم الخلايا العصبية المتصلة بما يلي من خلال الروابط الموزونة ويتم تمثيل كل خلية في دائرة وتأخذ رقمًا طبيعيًا (من 1:11) حيث يتم تمثيل الروابطبالسهام ويأخذij w حيث يشير الدليل إلى أنني أشير إلى رقم العقدة التي ينشأ منها السهم ، ويشير دليل j إلى رقم العقدة التي ينتهي عندها.يتم تنظيمها في طبقات بحيث يتم توصيل كل خلية في طبقة بجميع خلايا الطبقات السابقة واللاحقة .تبدأ الشبكة بطبقة الإدخال (1: 6) ، حيث تتوافق عقدتها مع أحد المتغيرات المستقلة ، وترتبط كل عقدة في طبقة الإدخال بجميع عقد الطبقة المخفية من (3-10).الطبقات في الطبقة الخارجة من الطبقة رقم (11) وهي عقدة واحدة أو أكثر تمثل المتغير التابع وهي نقطة التقاء الأسهم الخارجة من آخر طبقة مخفية.



شكل رقم (10) يوضح الشبكات العصبية

يتم حساب الوزن فيwijمن خلال مجموع ناتج الأوزون الذي يدخل العقدة التي ينشأ منها في قيم العقد التي ينطلق منها هذا الأوزون * wn = Wij .قيمة العقدةwn + يمكن اعتبار كل عقدة متغيرًا مستقلاً (1-6) أو كمجموعة (تفاعل) من المتغيرات المستقلة (7:11) ، والعقد 11 هي مجموعة غير خطية من القيم في العقد من 1: 6 بسببوجود وظيفة التنشيط إذا كانت وظيفة التنشيط خطية ولا توجد طبقة مخفية ، يتم تقليل الشبكة العصبية إلى انحدار خطي ، بينما يتم تقليل الشبكة العصبية إلى انحدار لوجستي تحت وظائف التنشيط غير الخطية ذات شكل معين.

في الشكل 11.يوضح الدقة العدد النسبي للأمثلة المصنفة بشكل صحيح أو بعبارة أخرى النسبة المئوية للتنبؤات الصحيحة في الشبكات العصبية.

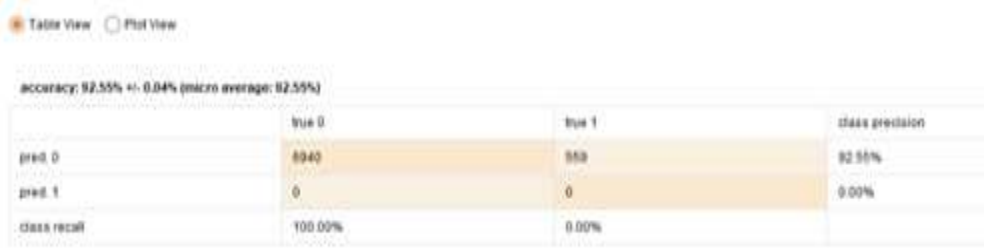


Table View Plot View

accuracy: 92.55% +/- 0.04% (micro average: 92.55%)

	true 0	true 1	class precision
pred. 0	8940	553	92.55%
pred. 1	0	0	0.00%
class recall	100.00%	0.00%	

الشكل رقم (11) يوضح دقة الشبكة العصبية , المصدر مستشفي الذرة الخرطوم إدارة الاحصاء(6).

12.الشكل يوضح خوارزمية العنقدة (*k-mean) يوضح عدد العناقد حيث العنقود 3583 =0عنصر و العنقود 487=1 عنصر و العنقود =2 1300 عنصر و العنقود 665 =3 عنصر و العنقود 1464 =4 عنصر .


Cluster Model

```
Cluster 0: 3583 items
Cluster 1: 487 items
Cluster 2: 1300 items
Cluster 3: 665 items
Cluster 4: 1464 items
Total number of items: 7499
```

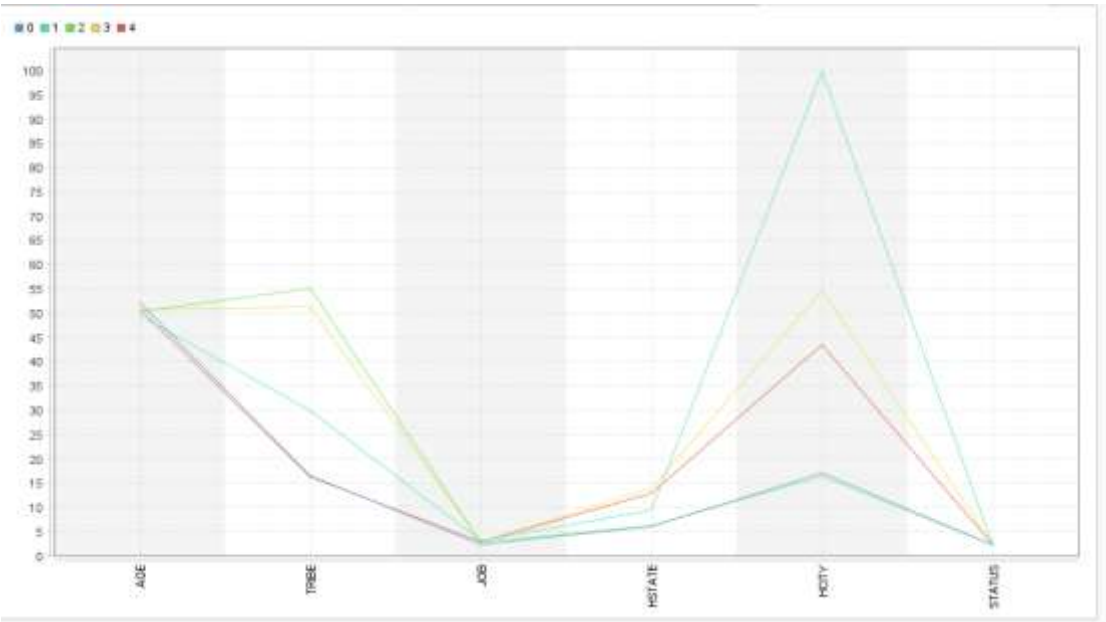
الشكل رقم (12)يوضح خوارزمية العنقدة (*k-mean) , المصدر مستشفي الذرة الخرطوم إدارة الاحصاء(6)

الشكل 13 يوضح نسب توزيع العنقود علي خصائص البيانات :

Attribute	cluster_0	cluster_1	cluster_2	cluster_3	cluster_4
AGE	52.225	49.468	50.292	50.704	50.897
TRIBE	16.484	29.928	55.061	51.262	16.249
JOB	2.307	3.080	2.763	2.884	2.848
HSTATE	6.032	9.396	6.201	13.794	12.849
HCITY	17.063	99.678	16.363	54.681	43.454
STATUS	2.230	2.230	2.203	2.265	2.253

الشكل رقم (13)يوضح نسب توزيع العنقود , المصدر مستشفي الذرة الخرطوم إدارة الاحصاء(6).

الشكل 14 توزيع الحقول فيما يتعلق بالعناقيد في شكل رسم بياني , حيث كان العنقود 0 اكثر توزيعا و ذلك حسب توزيع الولايات.



الشكل رقم(14)توزيع العناقيد , المصدر مستشفي الذرة الخرطوم إدارة الاحصاء.[6]

13. الجدول يوضع دقة خوارزميات التصنيف

الجدول رقم (2) دقة الخوارزميات

Algorithm	Accuracy
Neural Networks	92.55%
Naïve Bayes	99.16%
Decision Tree	92.53%
Rule model	94.40

النتائج والتوصيات

النتائج

إن نتائج نماذج تنقيب البيانات والنموذج الموضح بالشكل 7 والشكل 8 يحتوى مجموعة قواعد Rules من النوع (If...Then....)وهذا النوع من القواعد يسهل برمجته وإضامة لأي نظام طبي أو نظام معلومات وبالتالي يسهل على مبرمج وخبير بالحاسوب استخدام هذه القواعد المستنبطة والمستنتجة من كم هائل من البيانات والمعلومات. وفيما يلي ملخص هذه النتائج

1. تم تطبيق خوارمية شجرة القرار حيث كانت دقة الخوارزمية 92.53%.
2. معظم المصابين من الاناث ينتمون لولاية الخرطوم حيث بلغ عدد المصابين 1944 حالة من اجمالي الحالات والسبب الرئيسي في ذلك كثرة ابراج شبكة الاتصالات والمناطق الصناعية .
3. الاستنتاج بأن الفئات العمرية الأكثر عرضة للإصابة بسرطان الثدي هي ما بين (37-46) سنة .
4. إذا كان العمر اكبر من 16 اقل من 83 و الوظيفة مهنس و الحالة الاجتماعية متزوج فإن المصاب أنثى.
5. اذا كان العمر اقل من او يساوي 16 و الوظيفة ربة منزل فإن المصاب أنثي .
6. نتيجة شجرة القرارإذا كان المريض موظفاً ويقيم بالخرطوم وكان متزوجاً وعمره يزيد عن 45 سنة فإن المريض أنثى.
7. ان الاعوام اكثر إنتشارا للمرض حيث ان العام 2021 الاكثر انتشارا بعدد 3736 اصابة و العام الاقل 2012 بعدد 138 اصابة.

التوصيات

1- التوعية والإرشاد للكشف المبكر عن هذا المرض وادخال بيانات الكشف السريري والمخبري وألشعة
2- استخدام خوارزميات أخرى ومقارنة نتائجها بنتائج هذا البحث
3- تطبيق خوارزميات قواعد الارتباط لتحديد العلاقة بين المرضى
4- محاولة تعميم كافة الأمراض السرطانية بالمستشفى .

الخاتمة

بعد مراجعة نتائج البحث ظهرت أهمية بيانات مرضى سرطان الثدي وأهميتها في عمل إحصائيات عن عدد المرضى وأعمارهم وتاريخ المرض والمناطق التي ينتشر فيها المرض والاستفادة منها. الحد من انتشار المرض ومساعدة الأطباء في تشخيص المرض واتخاذ القرارات. هناك حاجة للتنقيب عن البيانات الخاصة بأمراض السرطان للاستفادة منها في اتخاذ القرار ، حيث أن إجراء العديد من الأبحاث في هذا المجال يمكن المؤسسات الصحية من وضع الخطة المتبعة وزيادة الكفاءة ، حيث اتضح من خلال الأبحاث والدراسات السابقة أن التنقيب عن البيانات هي إحدى الطرق الحديثة وذات الكفاءة العالية في هذا الميدان.

المراجع و المصادر

العلاق، ب. ع (2005) .الإدارة الرقمية: المجالات و التطبيقات. أبوظبي: مركز الإمارات للدراسات و البحوث الإستراتيجية،

العلي، ع.، قنديلجي، ع. إ.، العمري، غ (2006) .المدخل إلى إدارة المعرفة. الطبعة الأولى. عمان: دار المسيرة للنشر و التوزيع و الطباعة .

هبه عبدالله عبدالوهاب احمد. تقنيات التنقيب عن البيانات في الحقل الطبي (دراسة حالة الفشل الكلوي).

مستشفي الذرة الخرطوم سجلات إدارة الاحصاء

Bazsalica, M. & Naim, P. (2001) *Data mining pour le Web* . Paris: Eyrolles, p. 61.

Berry, J. A. M. &Linoff, G. S. (2004). *Data Mining Techniques For Marketing, Sales, and Customer Relationship Management* . 2nd ed. Indianapolis: Wiley Publishing, INC, p. 10.

Hand, D., Mannila, H., & Smyth, R. (2001). *Principles of Data Mining* . London: MIT Press, p. 1..

https://www.nejm.org/ -5