إستخدام تقنيات تنقيب البيانات في إستكشاف سرطان الثدي
(دراسة حاله مستشفي الذرة - الخرطوم)
(2021 - 2010)

أ.مرشد إبراهيم طالب مصطفى

محاضر-كلية علوم الحاسوب وتقانات المعلومات
جامعة القرآن الكريم والعلوم الإسلامية

المستخلص:

يهدف هذا البحث الي استخدام تقنيات التنقيب عن البيانات لإستكشاف المعرفة من سجلات المرضى ومعرفة اكثر الاعمار اصابة بالمرض لإجراء فحوصات وقائية مبكرة 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 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 , و 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. تم تطبيق خوارمية شجرة القرار حيث كانت دقة الخوارزمية 92.53% و 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,الاستنتاج بأن الفئات العمرية الأكثر عرضة للإصابة بسرطان الثدي هي ما بين (37-46) سنة.

الكلمات المفتاحية: البيانات ,المعلومات ,المعرفة ,مستودعات البيانات, التنقيب

The use of data mining techniques in the detection of breast cancer Case Study (Alzara Hospital-Khartoum) 2010- 2021

Morshed Ibrahim Talib Mustafa

Abstract:

This research aims to use data mining techniques to explore knowledge from patients' records and to know the ages most affected by the disease in order to conduct early preventive examinations of the 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 concerned with the use of techniques that extract and discover useful and independent knowledge through a large group of data. It helps in exploring hidden knowledge and unexpected models, in addition to exploring new rules and relationships that exist in rules. Large data helps to know the most causes of the spread of breast cancer and work to know the best methods of prevention and reduce the spread of the disease, which leads to a reduction in the mortality rate for patients, and the research follows the descriptive, analytical and experimental approach, where data and information related to monitoring records are collected, prepared, classified and tabulated. Then display and analyze it, and then adopt the constructivist approach to build a model capable of discovery in an effective way. The decision tree algorithm was applied, where the accuracy of the algorithm was 92.53%, and most of the patients were females belonging to the state of Khartoum, where the number of infected people reached 1944 cases out of the total number of cases, and the main reason for this is the large number of communication network towers and industrial areas, the conclusion is that the age groups most susceptible to breast cancer are what Between (3746-) years. Keywords: data, information, knowledge, data warehouses, prospecting .

أولاً : الإطار المنهجي :

المقدمة:

يتميز عصرنا الراهن (عصر الإنترنت والاقتصاد الرقمي) بالسيل العظيم والانتشار الواسع النطاق للبيانات حتى أضحى من المستحيل على المحللين استخلاص معلومات ذات معنى باللجوء فقط إلى المداخل التقليدية للتحليل التمهيدي للبيانات .مع وجود كميات كبيرة من البيانات

المخزنة في قواعد البيانات ومخازن البيانات ازدادت الحاجة إلى تطوير أدوات تمتاز بالقوة لتحليل البيانات واستخراج المعلومات والمعارف منها , ظهر ما يسمى بالتنقيب في البيانات كتقنية تهدف إلى استخراج المعرفة من كميات هائلة من البيانات وهي تقنية حديثة فرضت نفسها بقوة في عصر المعلوماتية واستخدامها يوفر للشركات والمنظمات في جميع المجالات القدرة على استكشاف والتركيز على أهم المعلومات في قواعد البيانات كما تركز تقنيات التنقيب على بناء التنبؤات المستقبلية واستكشاف السلوك والاتجاهات مما يسمح باتخاذ القرارات الصحيحة واتخاذها في الوقت المناسب.والتي تعتبر بدورها مرحلة من مراحل عملية أكثر تعقيدا هي استكشاف المعرفة في قواعد البيانات. حيث أن الشركات والمنظمات الرائدة اليوم تستخدم عملية استكشاف المعرفة في قواعد البيانات بشكل منهجي ومنظم بوصفها تشكل جوهر العمل الذي يعتمد عليه في تفعيل النشاط وتحقيق الميزة التنبؤية, يتناول هذا البحث الدراسة الاستكشافية والتنبؤية لسرطان الثدي بستخدام كميات كبير من البيانات من أجل دراسة هذا المرض الخطير جدا الذي يؤدي الي انهاء حياة كثير من المرضى .

مشكلة البحث:

لصعوبة 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 بألوسائل الاحصائية التقليدية.
2. صعوبة التنبوء بإحصائيات المرض في المستقبل.
3. عدم القدرة على الاستفادة القصوى من بيانات مريض سرطان الثدي

أهداف البحث:-

1. استخدام تقنيات التنقيب عن البيانات لإستكشاف المعرفة من سجلات المرضى.
2. معرفة اكثر الاعمار اصابة بالمرض لإجراء فحوصات وقائية مبكرة من المرض.
3. توفير نتائج تساهم في تقليل انتشار سرطان الثدي في الاعوام القادمة.
4. إستخدام تقنيات تنقيب البيانات الحديثة التي تعمل علي تسهيل تحليل البيانات .
5. استخدام **خوازمية** التصنيف وهي شجرة القرار((Tree Decision .

أسئلة البحث:

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

2- ما هي أكثر الولايات الاكثر انتشارا للمرض؟

3- ماهي القواعد التي توضع انتشار المرض؟

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 .**

حدود البحث:-

- **الحدود المكانية:مستشفى الذرة بالخرطوم .**
- **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.

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.

ثانياً الدراسات السابقة:

1.الدراسة الاولي: مدثر يونس حسن إبراهيم (2018):

التنبؤ بمستوى 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 (2017):

استخدام تقنيات التنقيب عن البيانات لمرضى الفشل الكلوي(دراسة حالة مستشفى احمد قاسم)

الملخص:

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):

استخدام التنقيب عن البيانات لبناء خطط علاج لمرضى السكر

الملخص:

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 ، Roc 0.644، 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):

استخدام التجميع والتصنيف للتنبؤ بانتشار مرض التهاب الكبد الوبائي ، دراسة حالة (ولاية الخرطوم)

الملخص:

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:

تقنيات استخراج البيانات في المجال الطبي (دراسة حالة الفشل الكلوي

الملخص:

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

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.

مقارنة الدراسات السابقة:

1. مقارنة دراستنا بالدراسة الأولى لأن دراستنا ركزت على Crispmethodology وأداة Rapid Miner 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 Rapid Miner 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 Rapid Miner 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 Rapid Miner 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

ثانيا: الإطار النظري لتنقيب البيانات:

1 -مقدم التنقيب عن البيانات :

مع وجود كميات هائله من البيانات المخزنه في قواعد البيانات الضخمه ازدادت الحاجه إلى تطوير أدوات تمتاز بالدقه لتحليل البيانات واستخراج المعلومات والمعارف منها ,ومن هنا ظهر ما يسمي بالتنقيب عن البيانات كتقنيه تهدف الى استنتاج المعرفه من كميات هائله من البيانات ,ولإهميه هذا العلم تم استخدامه في المجال الطبي وفي تشخيص الامراض التي يصعب تشخيصها , أدى الانتشار الواسع لتقنية المعلومات وسهولة إتاحتها إلى تضخم حجم المعلومات بصورة استباقية لم يشهدها التاريخ من قبل، مما جعل من قضية البيانات الضخمة على الإنترنت مثاراً للجدل، من حيث جدوى وجودها بهذه الصورة العشوائية. وعندما نتحدث عن البيانات الضخمة، فإننا

نتحدث عن كميات لا يمكن تخيلها من البيانات متعددة الأنواع والمصادر بحجم يصل إلى المئات من التيرابايت أو حتى البيتابايت ذلك أدى إلى ازدياد الحاجة إلى تطوير أدوات تمتاز بالقوة لتحليل البيانات واستخراج المعلومات والمعارف منها، فالأساليب التقليدية والإحصائية لا تستطيع أن تتعامل مع هذا الكم من الهائل لذا تستخدم أدوات ذكية لمعالجة هذه البيانات.

من هنا ظهر ما يسمى باستخراج البيانات Data Mining كتقنية تهدف إلى استنتاج المعرفة من كميات هائلة من البيانات، تعتمد على الخوارزميات الرياضية والتي تعتبر أساس التنقيب عن البيانات وهي مستمدة من العديد من العلوم مثل علم الإحصاء والرياضيات والمنطق وعلم التعلم، والذكاء الاصطناعي والنظم الخبيرة، وعلم التعرف على الأنماط ،وعلم الآلة. وغيرها من العلوم والتي تعتبر من العلوم الذكية وغير التقليدية.

ظهر التنقيب في البيانات (Data mining) في أواخر الثمانيات وأثبت وجوده كأحد الحلول الناجحة لتحليل كميات ضخمة من البيانات، وذلك بتحويلها من مجرد معلومات متراكمة وغير مفهومة (بيانات) إلى معلومات قيِّمة يمكن استغلالها و الاستفادة منها بعد ذلك.وقد اجتذبت مرحلة التنقيب في البيانات الكثير من الاهتمام في الأوساط البحثية على مدي العقد الماضي، في محاولة لتطوير خوارزميات قابلة للتوسع والتكيف مع كميات متزايدة من البيانات في البحث عن أنماط معرفية ذات معنى. وقد نمت حزم من الخوارزميات والبرمجيات و بشكل كبير خلال العقد الماضي، إلى حد أن التوسع قد جعل من الصعب على العاملين في هذا الحقل تتبع التقنيات المتاحة لحل مهمة معينة.

التنقيب عن البيانات (أحيانا تسمى إكتشاف المعرفة) هي عملية تحليل البيانات من منظورات مختلفة واستخلاص علاقات بينها وتلخيصها إلى معلومات مفيدة, مثل معلومات يمكن أن تسهم في زيادة الربح، تخفيض التكاليف، أَو كليهما معا.تقنيا، يعتبر التنقيب عن البيانات عملية لإيجاد الإرتباطات بين العشرات من الحقول في قواعد البيانات العلائقية الكبيرة.

2-المصطلحات المستخدمة في البحث البيانات والمعلومات والمعرفة ومستودعات البيانات:

- **البيانات Data:** هي عبارة عن الحقائق والأرقام والنصوص التي يمكن أن تعالج من قبل الحاسب.
- - **المعلومات Information:** النماذج والعلاقات بين تلك البيانات والتي تشكل معلومات مفيدة.
- **المعرفة Knowledge:** المعلومات السابقة يمكن أن تحول إلى معرفة حول الأنماط التاريخية أو التوقعات المستقبلية، مثال معلومات عن حركة المبيعات والمشتريات للزبائن يمكن أن تزودنا بمعرفة عن سلوكهم الشرائي، فيساعدنا ذلك في معرفة أي من المواد تحتاج إلى ترويج أكثر. (4)
- المستخدمة في التحليلات الزمنية واكتشاف المعرفة واتخاذ القرارات, فهي مصممة

خصيصا لاستخلاص البيانات ومعاجتها وتمثيلها وتقديمها بصورة مناسبة لهذه الأغراض، وتخزن كمية ضخمة من البيانات قد تكون من مصادر مختلفة، مثلا عدة قواعد بيانات من عدة نماذج. (4)

3 -بماذا يمكن أن نستخدم التنقيب عن البيانات ؟

على فرض أنك تملك متجرا كبيرا يحتوي هذا المتجر على عدد كبير من السلع المختلفة، وهناك عوامل كثيرة تؤثر على عملك، منها "عوامل داخلية" مثل السلع و الأسعار ومهارات الباعة، و"عوامل خارجية" مثل وضع الزبون والمنافسة والمؤشرات الإقتصادية. ففي حال أردت الإستعلام عن منتج معين و تربط هذا الإستعلام بالعوامل الداخلية والخارجية فإنك تحتاج إلى التنقيب عن البياناتData Mining للحصول على نتيجة جيدة. (4)

4 -أمثلة عن التنقيب عن البيانات:

في إحدى المتاجر الكبيرة حيث يحتوي هذا المتجر على تنوع كبير من الأطعمة لاحظ الفريق المهتم بالزبائن أن معظم الزبائن الذي يشترون الحليب يشترون الخبز معه مما يمكن التاجر من إعادة ترتيب الأطعمة في المتجر وفقا لما يراه مناسب لزيادة أرباح المتجر, مثلا بوضع الخبز بجانب الحليب. ليكن لدينا سلسلة من المطاعم وليكن لدينا زبائن يأخذون وجبة بشكل نموذجي, هنا يمكن ان ننقب بيانات شراء الزبائن لتحديد ماهي الوجبة المطلوبة. بالتنقيب في بيانات متجر لبيع لوازم السفر والرحلات, وجد أن من يشتري أكياس نوم وأحذية سفر وخيمة فسيقوم أيضاً بشراء حقيبة ظهر للسفر.

5 -مراحل اكتشاف المعرفة في التنقيب عن البيانات:

1. اختيار البيانات إنها الخطوة الموجهة نحو تحديد مصدر البيانات في الدراسة ، بما في ذلك استخدام البيانات الخارجية العامة ، وهي مرحلة يتم فيها تحديد البيانات المناسبة واسترجاعها من قاعدة البيانات.
2. تهيئة البيانات هي مرحلة معالجة وعزل البيانات المهمة أو المفقودة أو المحتوية على البيانات المتبقية مثل الإلغاء ، المعلومات المتكررة ، التصحيح الرسمي ، معالجة البيانات المفقودة وجعلها جاهزة للتطبيق. وتشمل هذه: المرحلة (تنظيف البيانات ، حذف البيانات المفقودة ، اشتقاق البيانات ، دمج البيانات
3. تحويل البيانات هي عملية نقل البيانات المحددة إلى نموذج مناسب للخوارزميات والتطبيقات التي سيتم استخدامها في البحث قد تتطلب بعض الخوارزميات وجود بيانات بتنسيق معين قبل التطبيق
4. التنقيب عن البيانات في هذه المرحلة ، سيتم تطبيق طريقة ذكية لاستخراج النماذج المفيدة قدر الإمكان.
5. تقييم الأنماط بعد استخراج النماذج المهمة التي تمثل المعرفة ، يتم تقييمها بناءً على مقاييس محددة في بيئة المشكلة.

6. تمثيل المعرفة إنها المرحلة الأخيرة من اكتشاف المعرفة في قواعد البيانات ، والتي يراها المستفيد ، وهي المرحلة الأساسية التي تستخدم الأسلوب البصري لمساعدة المستفيد على فهم وتفسير النتائج المستخرجة. (4)

7 - مراحل عملية التنقيب عن البيانات:

1. فهم طبيعة العمل الشرط الأول لاكتشاف المعرفة هو فهم المشاكل والقضايا التي يواجهها العمل. بمعنى آخر ، كيفية تحقيق أكبر فائدة من التنقيب في البيانات ، الأمر الذي يتطلب صيغة واضحة ومحددة لأهداف العمل.
2. فهم البيانات تعد مسألة معرفة طبيعة وطبيعة البيانات عاملاً مهمًا في نجاح التنقيب عن البيانات واكتشافها. إن معرفة البيانات جيدًا يعني مساعدة المصممين على استخدام الخوارزميات أو الأدوات المستخدمة في قضايا محددة بدقة عالية. وهذا يؤدي إلى تعظيم فرص النجاح بالإضافة إلى الزيادة فاعلية وكفاءة نظام اكتشاف المعرفة. لا يحتاج التنقيب عن البيانات إلى جمع البيانات في مستودع البيانات ، ولكن إذا كان مستودع البيانات موجودًا في المؤسسة ، فمن الأفضل عدم احتكار المستودع مباشرةً لغرض التنقيب عن البيانات. (4)

8 -تطبيق خوازمية التصنيف وهي شجرة القرار(Tree Decision):

1 -مجموعة البيانات وخصائص الدراسة وانواعها:

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

الجدول رقم (1) يوضح خصائص البيانات

2 - شجرة القرار: يوضح الشكل شكل شجرة القرار والعلاقات بين الحقول:

الشكل رقم (1) يوضح شجرة القرار المصدر برنامج (Rapid Miner)

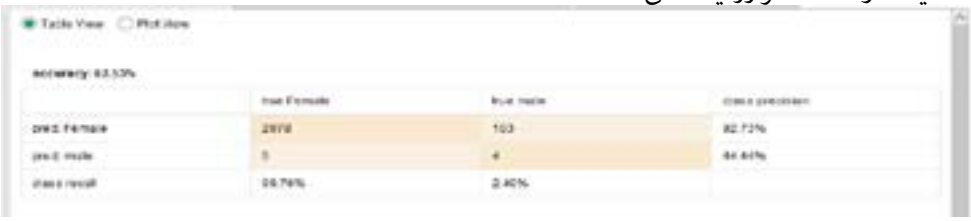
3 -يوضح الشكل نسب ارتباطات شجرة القرار بين الحقول ونسبة توزيعها على كل فئة

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 (2) يوضح وصف قواعد شجرة القرار المصدر برنامج (Rapid Miner)

4 -العـدد النسـبي للأمثلـة المصنفـة بشـكل صحيـح أو بعبـارة أخـرى النسـبة المئويـة للتنبـؤات الصحيحـة و دقـة الخوارزميـة هـي92.53%..



الشكل رقم (3) يوضح النسبة المئوية للتنبؤات الصحيحة.المصدر برنامج (Rapid Miner)

ثالثا: النتائج :

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.

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 .

التوصيات:

1. إنشاء مستودع بيانات كامل لتوفير جميع المعلومات والبيانات التي يحتاجها المحللون لمساعدتهم في التنقيب عن البيانات واكتشاف المعرفة
2. التوعية والإرشاد للكشف المبكر عن هذا المرض
3. استخدام خوارزميات أخرى ومقارنة نتائجها بنتائج هذا البحث
4. تطبيق خوارزميات قواعد الارتباط لتحديد العلاقة بين المرض
5. محاولة تعميم كافة الأمراض السرطانية بالمستشفى .

الهوامش:

(1) بشـير عبـاس، العـلاق، الإدارة الرقميـة: المجـالات و التطبيقـات، مركـز الإمـارات للدراسـات و البحـوث الإسـتراتيجية، ابوظبـي، 2005، ص 84.

(2) Bazsalica M., Naim P., Data mining pour le Web, éd. Eyrolles, Paris, 2001, P. 61.

(3) عبـد السـتار العـلي، عامـر إبراهيـم قنديلجـي، غسـان العمـري، المدخـل إلى إدارة المعرفـة،دار المسـيرة للنـشر و التوزيـع و الطباعـة، الطبعـة الأولى، عـمان، 2006، ص. 157.

(4) تقنيات التنقيب عن البيانات في الحقل الطبي (دراسه حالة الفشل الكلوي.

(5) Hand d., Mannila H., Smyth R., Principles of Data Mining, MIT Press, London, 2001, p. 01.

(6) Bazsalica M., Naim P., Op. Cit., pp.6869-.

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

(8) »C:\Program Files\RapidMiner\RapidMiner Studio\RapidMiner Studio.exe«

(9) مستشفي الذرة الخرطوم إدارة الاحصاء.

(10) هبه عبدالله عبدالوهاب احمد (2020) زمن الدخول) 10-9-2021

(11) : https://www.nejm.org/