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

إستخدام(Rapid Miner) في تحليل بيانات بسرطان الثدي بالسودان: دراسة حاله مستشفي الذرة – الخرطوم(2010-2021م)

طارق عبدالكريم[1] و مرشد ابراهيم[2]

1جامعة النيلين

2جامعة القران الكريم

المؤلف:morshedsadua@gmail.com

تاريخ الاستلام: 31 يوليو 2025م تاريخ القبول: 19 نوفمبر 2025م

المستخلص:

يهدف البحث الي إستخدام الطرق الإحصائية في برامج (Rapid Miner) في إستخراج النتائج الإحصائية وتحديد نسبة الإصابة بالمرض وفئات الفئات العمرية الأكثر عرضة لهذا المرض من خلال عملية التنقيب عن بيانات المرضى ومعرفة اكثر الاعمار اصابة بالمرض لإجراء فحوصات وقائيةمبكرة من المرض وهتمة الدراسة علي العمل علي تحليل دقيق لكمية كبيرة من البيانات المتوفرة لعدد من السنوات وإستخدام الطرق الإحصائية التي تساعد علي اتخاذ قرارات تساعد في المعرفة بمعدلات انتشار المرض في المستقبل وتوفير البيانات الازمة التي تساعد علي ارشادات ونصائح في افضل الطرق لتجنب انتشار سرطان الثدي.ويتبع البحث المنهج الوصفى التحليلى والتجريبي ، حيث يتم جمع البيانات والمعلومات الخاصه بسجلات المراقبة وإعدادها وتصنيفها وتبويبها ومن ثم عرضها وتحليلها, وتم التوصل للنتائج من خلال إستخدام الطرق الإحصائية في برنامج (Rapid Miner) تم التوصل الي ان ولاية الخرطوم الاكثر إصابة بعدد 1944 اصابة بين ان ولاية النيل الازرق الاقل إصابة بعدد 45 اصابة و انالولاية الشمالية هي الاكثر اصابة مقارنتا بعدد السكان مع نسبة الاصابة.ان خوارزمية (Naïve Bayes) توضح العلاقة بين الأعمار وتوزيعها في الولايات فيNaïve Bayes , حيث كانت اكثر الولايات توزرعا ولاية الخرطوم. ان عدد الاصابات حسب المحليات كان العدد الاكبر في محلية الخرطوم بعدد 1230إصابة .

الكلمات المفتاحية:تنقيب بيانات -معرفة - الطرق الإحصائية- الوصفى التحليلى- جمع البيانات

Using (Rapid Miner) in Breast Cancer Data Analysis in Sudan: A Case Study of Al-Dora Hospital – Khartoum (2010-2021)

Tarig Abdelkarim[1] and Murshid Ibrahim[2]

1Elneelain University, Khartoum
2Holy Quraan University, Omdurman

Received: 31st July, 2025 Accepted: 19th November, 2025

Abstract:

The research aims to use statistical methods in (Rapid Miner) programs to extract statistical results and determine the incidence of the disease and the age groups most vulnerable to this disease through the process of excavating patient data and knowing the most affected ages for early preventive examinations of the disease. Careful analysis of a large amount of data available for a number of years and the use of statistical methods that help in making decisions that help in knowing the rates of disease prevalence in the future and providing the necessary data that helps guides and advice on the best ways to avoid the spread of breast cancer. The research follows the analytical and experimental descriptive approach, where data and information related to monitoring records are collected, prepared, classified and tabulated, and then displayed and analyzed, and the results were reached through the use of statistical methods in the (Rapid Miner) program. The Blue Nile State has the least infection with 45 infections, and the tribes with the most prevalence of the disease have 1025 infections, and the least is the Barqawi tribe with 1 infection.

Keywords*: data mining, knowledge, statistical methods, descriptive analysis, data collection*

أولاً : الإطار المنهجي :

المقدمة:

مع وجود كميات كبيرة من البيانات المخزنة في قواعد البيانات ومخازن البيانات ، زادت الحاجة إلى تطوير أدوات قوية لتحليل البيانات واستخراج المعلومات والمعرفة منها. من هنا ، ظهر ما يسمى بالتنقيب في البيانات كتقنية تهدف إلى استخراج المعرفة من كميات هائلة من البيانات وإيجاد علاقة منطقية تلخص البيانات. بطريقة جديدة مفهومة ومفيدة لصاحب البيانات ، هي تقنية حديثة فرضت نفسها بقوة في عصر المعلومات ، واستخدامها يوفر للدولة والشركات والمؤسسات والمستشفيات في جميع المجالات القدرة على استكشاف و التركيز على أهم المعلومات في قواعد البيانات ، وتركز تقنيات الاستكشاف على بناء التنبؤات. بعد أن طور العلماء أجهزة الكمبيوتر ، أدرك المجتمع والعالم كله أن هذه الأجهزة الجديدة ستوفر العديد والعديد من الخدمات للبشرية جمعاء ، خاصة في مجال المعلومات والتخزين والمعالجة والاسترجاع ، وبعد هذا التاريخ قبل عقد من الزمن ، الأطباء والمتخصصون وبدوره بدأ بمحاولة الاستفادة من هذه التقنيات بشكل حقيقي من خلال تطوير فكرة إدارة المعلومات ودور الحاسب الآلي في الطب والرعاية الصحية التي تعد من أهم المجالات العلمية وأكثرها انتشارا و مؤثر ، ولا يزال الطب يبحث عن مزيد من التطور باستخدام جميع وسائل العلم المتاحة ، وأهمها التكنولوجيا وأنظمة المعلومات وأدوات التنقيب عن البيانات لتحليل الأمراض ومدى انتشارها وطرق الحماية. تهتم العديد من الدول المتقدمة ببيانات المرضى أكثر من البيانات الأخرى لأنها تدرك تمامًا أهمية الصحة ، ونجد أن الهيكل الأساسي للدولة هو القوة العاملة ، ووجودها بصحة جيدة يعني مدى انتشار الصحة. الوعي ، وهذا بدوره يؤدي إلى تطوير جيل جديد سليم في المجتمع.

مشكلة البحث:

لصعوبة التعامل مع البيانات الكبيرة وصعوبة الوصول الي نتائج مع البيانات الكبيرة ويمكن توضيح مشكلة البحث في الاتي :

1. عدم استخدام أدوات التصنيف لوصف الفئات العمرية الأكثر عرضة للإصابة بمرض انتشار المرض.
2. صعوبة تحليل مجموعة كبيرة من البيانات بألوسائل الاحصائية التقليدية.

أهداف البحث:-

1. إستخدام الطرق الإحصائية في برامج (Rapid Miner) في إستخراج النتائج الإحصائية.
2. تحديد نسبة الإصابة بالمرض والولايات الأكثر عرضة لهذا المرض من خلال عملية التنقيب عن بيانات المرضى.
3. معرفة الولاية الاكثر إصابة بالمرض مقارتنا بعدد الاصابات مع نسبت السكان.
4. إستخدام خوارزمية (Naïve Bayes) التي توضح العلاقة بين الأعمار وتوزيعها في الولايات فيNaïve Bayes .

أسئلة البحث:

1 - ماهي السنوات الاكثر انتشارا للمرض ا؟
2- ما هي العلاقة بين الأعمار وتوزيعها في الولايات ؟
3-هل الولاية الاكثر إصابة بالمرض مقارتنا بعدد الاصابات مع نسبت السكان؟

اهمية البحث:

العمل علي تحليل دقيق لكمية كبيرة من البيانات المتوفرة لعدد من السنوات وإستخدام الطرق الإحصائية التي تساعد علي اتخاذ قرارات تساعد في المعرفة 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 الخرطوم.
- 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

الإطار النظري

الدراسات السابقة:

الدراسة الاولي: أم كلثوم صباحى محمد حمدون وأسراء فتحى يعقوب محمد على بعنوان: **إستخدام تقنية تنقيب البيانات فى أمراض السرطان (بالتطبيق على مركز الخرطوم للعلاج بالاشعه و الطب النووي)**. يهدف هذا البحث لحل إحدى المشاكل التى يعانى منها الأطباء و هى مشكل تشخيص أمراض السرطان التى تؤدى إلى الوفاة , كما أنه توجد بيانات ضخمة دون الإستفادة منها , لذا جاء هذا البحث لحل هذه المشكلة بالإضافة إلى مساعدة الأطباء لإتخاذ القرار الصحيح .

لقد قمنا بحمد الله بتحليل و تصميم قاعدة بيانات لحفظ سجلات المرضى، و من ثم القيام بعملية تنقيب البيانات من خلال أفضل الخوارزميات، حيث تم إستخدام تقنية Clementine و الدذى يتكون من واجهة واحدة تحتدوى على عدد من الأدوات و تمت عملية التنقيب من خلال خوارزمية التصنيف حيث إستخدمنا خوارزمية واحدة من خوارزميات التصنيف و هى خوارزمية شجرة القراراتDecision Treeبتطبيق مصنفC5.0 و بيانات أمراض و مرضي السرطان مدن قاعددة البياندات التى قمنا بتصدميمها ,و فى الخطوة الاخيرة قمنا ببناء النموذج عن طريق مصنفC5.0.واستخدمت الدراسة Waka.

الدراسة الثانية: عبد القوي بلاشيا بعنوان: تحليل للتنبؤ بمعدل البقاء على قيد الحياة لمرضى سرطان الثدى بإستخدام تقنيات التنقيب. نقدم تحليل للتنبؤ بمعدل البقاء على قيد الحياة لمرضى سرطان الثدى بإستخدام تقنيات التنقيب هم : شجرة القرار و بايز و الشبكات العصبية , كما تم إستخدام مجموعة بيانات تتألف من 151776 من السلات و الحقول متاح منها 16 حقل فقط من قاعدة البيانات , أجريت تجارب عدة بإستخدام هذه الخوارزميات التى حققت عرض متشابه للتنبؤ . وجد أن شجرة القرار لديها أداء أفضل بكثير من التقنيات الأخرى .

الدراسة الثالثة: طريقة تعتمد على تقنيات التنقيب عن البيانات لتحليل تكرار سرطان الثدي(2020)

السرطان مرض يتطور باستمرار ، ويؤثر على عدد كبير من الناس في جميع أنحاء العالم مستوى البحث لتطوير أدوات تعتمد على تقنيات التنقيب عن البيانات التي تسمح باكتشاف سرطان الثدي أو الوقاية منه.تلعب الأحجام الكبيرة من البيانات دورًا أساسيًا وفقًا للأدبيات التي تم الرجوع إليها ، وقد تم إنشاء مجموعة كبيرة ومتنوعة من مجموعات البيانات الموجهة لتحليل المرض ، وفي هذا البحث تم استخدام مجموعة بيانات سرطان الثدي ، والغرض من البحث المقترح هو تقديممقارنة بين خوارزميات تصنيف J48 وrandomforestوNaiveBayesو NaiveBayes Simpleو SMO Poli-kernel و SMO RBF-Kernel ، المدمجة مع خوارزمية مجموعة K-Meansالبسيطة لإنشاء نموذج يسمح بالتصنيف الناجح للمرضى الذين هم أو غير -عودة سرطان الثدي بعد الخضوع لعملية جراحية سابقة لعلاج المرض المذكور ، وأخيراً الطرق التي حصلت على أفضل المستويات هي98.5 SMO Poly-Kernel + Simple K-Means ٪من الدقة ، 98.5٪ استدعاء ، 98.5٪ TPRATEو 0.2٪ .FPRATEتشير النتائج التي تم الحصول عليها إلى إمكانية استخدام أدوات حسابية ذكية تعتمد على طرق استخراج البيانات للكشف عن تكرار الإصابة بسرطان الثدي لدى المرضى الذين خضعوا لعملية جراحية سابقًا.

المنهجية:

لتطوير البحث المقترح ، نبدأ في البداية بالحصول على مجموعة البيانات المسماة Breast Cancer Wisconsin المأخوذة من [17] ، والتي كان من الضروري خلالها إجراء مرحلة المعالجة المسبقة للبيانات المسماة المرحلة رقم 1 حيث نسلط الضوء على تحقيق تحليلموازنة البيانات ، لاحقًا في المرحلة الثانية ، عملية التدريب واختبار طرق التصنيف المستخدمة حيث تمت مقارنة خوارزميات DT و NB و SVM من خلال

مقاييس الدقة وتقييم التغطية والمعدل الإيجابي الحقيقي والمعدل الإيجابي الخاطئ ، وأخيراً فيفي المرحلة 3 ، يتم أخذ أفضل طريقة تصنيف باستخدام الوسيلة العنقودية البسيطة ، ويتم مقارنة النتيجة التي تم الحصول عليها فيما يتعلق بالنتائج التي تم الحصول عليها من خلال التصنيف فقط.تم إجراء عملية التجريب باستخدام أداة التنقيب عن البيانات WEKA .

الدراسة الرابعة :تطبيق التنقيب عن البيانات لتحليل بيانات سرطان الثدي(2015). التنقيب عن البيانات ، المعروف أيضًا باسم اكتشاف المعرفة في قواعد البيانات (KDD) هو عملية البحث تلقائيًا عن كميات كبيرة من البيانات عن الأنماط.على سبيل المثال ، قد يشير النمط السريري إلى أن الأنثى المصابة بالسكري أو ارتفاع ضغط الدم تكون أسهل في المعاناة من السكتة الدماغية لمدة 5 سنوات في المستقبل.بعد ذلك ، يمكن للطبيب تعلم معرفة قيمة من عمليات التنقيب عن البيانات.هنا ، نقدم دراسة تركز على التحقيق في تطبيق تقنيات الذكاء الاصطناعي واستخراج البيانات على نماذج التنبؤ بسرطان الثدي.تم استخدام الشبكة العصبية الاصطناعية ، وشجرة القرار ، والانحدار اللوجستي ، والخوارزمية الجينية للدراسات المقارنة ، كما تم استخدام الدقة والقيمة التنبؤية الإيجابية لكل خوارزمية كمؤشرات للتقييم.تم الحصول على 699 سجلًا من مرضى سرطان الثدي في جامعة ويسكونسن ، وتم دمج تسعة متغيرات توقع ، ومتغير نتيجة واحد لتحليل البيانات متبوعًا بالتحقق من الصحة بعشرة أضعاف.أظهرت النتائج أن دقة نموذج الانحدار اللوجستي كانت 0.9434 (الحساسية 0.9716 والنوعية 0.9482) ، ونموذج شجرة القرار 0.9434 (الحساسية 0.9615 ، والنوعية 0.9105) ، ونموذج الشبكة العصبية 0.9502 (الحساسية 0.9628 ، والنوعية 0.9273) ، والخوارزمية الجينية.موديل 0.9878 (حساسية 1 ، خصوصية 0.9802).(كانت دقة الخوارزمية الجينية أعلى بكثير من متوسط الدقة المتوقعة عند 0.9612.كانت النتيجة المتوقعة لنموذج الانحدار اللوجستي أعلى من تلك الخاصة بنموذج الشبكة العصبية ولكن لم يلاحظ أي فرق كبير.كان متوسط الدقة المتوقعة لنموذج شجرة القرار 0.9435 وهو أدنى مستوى من جميع النماذج التنبؤية الأربعة.كان الانحراف المعياري للتحقق المتقاطع من عشرة أضعاف غير موثوق به إلى حد ما.أوضحت هذه الدراسة أن نموذج الخوارزمية الجينية حقق نتائج أفضل من نماذج التنقيب عن البيانات الأخرى لتحليل بيانات مرضى سرطان الثدي من حيث الدقة الكلية لتصنيف المريض ، وتعبير وتعقيد قاعدة التصنيف.أظهرت النتائج أن الخوارزمية الجينية الموصوفة في الدراسة الحالية كانت قادرة على إنتاج نتائج دقيقة في تصنيف بيانات سرطان الثدي وأن قاعدة التصنيف التي تم تحديدها كانت أكثر قبولا وفهمًا.

الدراسة الخامسة: تطبيق تقنيات التنقيب عن البيانات للتنبؤبسرطان الثدي(2019). يعد سرطان الثدي من الأمراض التي تسبب عددًا أكبر من الوفيات خلال عام.يعتبر سرطان الثدي ثاني أكثر الأمراض المسببة للوفاة بين النساء، وفي كندا هو سبب رئيسي للوفاة.الاكتشاف المبكر لسرطان الثدي يجعله أكثر أنواع السرطان قابلية للشفاء من بين أنواع السرطان الأخرى، ويضمن الكشف المبكر والفحص الدقيق لسرطان الثدي زيادة معدل البقاء على قيد الحياة للمرضى.تتمتع تقنيات التنقيب عن البيانات بسمعة متزايدة في المجال الطبي بسبب القدرة التشخيصية العالية والتصنيف المفيد.يمكن أن تساعد أساليب التعلم الآلي الممارسين على تطوير أدوات تسمح باكتشاف المراحل المبكرة من سرطان الثدي .الهدف من هذه الدراسة هو التنبؤ بسرطان الثدي باستخدام-k الأقرب إلى الجار (KNN) ، آلة المتجهات الداعمة (SVM)، الغابة العشوائية.(RF) علاوة على ذلك ، أجرينا مقارنة تفصيلية بين الطرق الثلاث.يمكن استخدام جميع الأساليب بمفردها أو مع التعلم الجماعي لبناء مصنف أكثر تعقيدًا.نستخدم مجموعة بيانات سرطان الثدي في ولاية ويسكونسن لتدريب جميع المصنفات والتحقق من صحتها.ثم يتم قياس مصفوفة الأداء، أي الدقة والتذكر والدقة في مجموعة بيانات تدريب واختبار مختلفة.تظهر طريقة التعلم الجماعي القائمة على الحد الأقصى للتصويت أعلى دقة (98.9٪) مقارنة بتقنيات التصنيف الأخرى.

الدراسة الرابعة عشر: حمزة سعد(2020). **تقنيات التنقيب عن البيانات في التنبؤبسرطان الثدي**. سرطان الثدي، الذي يمثل 23٪ من جميع أنواع السرطان، يهدد مجتمعات البلدان النامية بسبب ضعف الوعي والعلاج.يساعد التشخيص المبكر كثيرًا في علاج المرض.أجريت الدراسة الحالية بهدف تحسين عملية التنبؤ واستخراج الأسباب الرئيسية التي أثرت على سرطان الثدي.المواد والطرق: تم جمع البيانات بناءً على ثماني سمات لـ 130 سيدة ليبية في المراحل السريرية المصابة بهذا المرض.تم استخدام التنقيب عن البيانات من خلال تطبيق ستة خوارزميات للتنبؤ بالمرض بناءً على المراحل السريرية.تكتسب جميع الخوارزميات دقة عالية ، لكن شجرة القرار توفر أعلى مخطط دقة لشجرة القرار المستخدمة لبناء القواعد من كل عقدة ورقية.متغيرات الترتيب المطبقة لاستخراج المتغيرات المهمة ودعم القواعد النهائية للتنبؤ بالمرض.النتائج: حصلت جميع الخوارزميات المطبقة على تنبؤ عالي وبدقة مختلفة.قدمت القواعد 1 و 3 و 4 و 5 و 9 مجموعة فرعية نقية ليتم تأكيدها كقواعد مهمة.ساهمت خمسة متغيرات إدخال فقط في بناء القواعد ، ولكن ليس لجميع المتغيرات تأثير كبير.الخلاصة: يلعب حجم الورم دورًا حيويًا في بناء جميع القواعد ذات التأثير الكبير.متغيرات الوراثة وجانب الثدي وحالة سن اليأس لها تأثير ضئيل في التحليل ، لكنهم قد ينظرون في نتائج ملحوظة باستخدام استراتيجية مختلفة لتحليل البيانات.

مقارنة الدراسات السابقة:

مقارنة دراستنا بالدراسة الاولي: لأن دراستنا ركزت على منهجية Crisp و استخدمت دراستنا أداة Rapid Miner واستخدام تقنيات استخراج البيانات (التجميع والتصنيف) واستخدام الخوارزميات ومقارنتها مع بعضها البعض و استخدمت هذه الدراسة إستخدام تقنية Clementine و توافقت مع دراستنا في استخدام خوارزمية شجرة القراراتDecision Tree, واختلفت في استخدام Weka.

مقارنة دراستنا بالدراسة الثانية: لأن دراستنا ركزت على منهجية Crisp و استخدمت دراستنا أداة MinerRapid واستخدمت الدراسة شجرة القرار و بايز و الشبكات العصبية و تشابهت مع دراستنا في استخدام بايز.

مقارنة دراستنا بالدراسة الثالثة حيث اتفقت مع دراستنا في دراسة سرطان الثدي و قامت بمقارنة بين خوارزميات تصنيف J48 وrandomforestوNaiveBayesو NaiveBayes Simple وSMO Poli-kernel و SMO RBF-Kernel ، المدمجة مع خوارزمية مجموعة K-Means البسيطة لإنشاء نموذج يسمح بالتصنيف الناجح للمرضى, واختلفت في استخدام Weka.

مقارنة دراستنا بالدراسة الرابعة: واتفقت مع دراستنا في إستخراج البيانات على نماذج التنبؤ بسرطان الثدي.تم استخدام الشبكة العصبية الاصطناعية ، وشجرة القرار ، والانحدار اللوجستي ، والخوارزمية الجينية و إختلفت في منهجية البحث حيث إستخدمت دراستنا منهجية Crisp.

مقارنة دراستنا بالدراسة الخامسة: حيثالهدف من هذه الدراسة هو التنبؤ بسرطان الثدي باستخدام-k الأقرب إلى الجار (KNN) ، آلة المتجهات الداعمة (SVM) ، الغابة العشوائية.(RF) علاوة على ذلك ، أجرينا مقارنة تفصيلية بين الطرق الثلاث, وإختلفت عن دراستنا في إدراة التنقيب .

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 قيّمة 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، أو كليهما معا.تقنيا، يعتبر التنقيب عن البيانات عملية لإيجاد الإرتباطات بين العشرات من الحقول في قواعد البيانات العلائقية الكبيرة.

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)

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)

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(2).

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)

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 .

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)

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. (2)

تنقيب البيانات والعلوم الأخرى:

يُعتبر تنقيب البيانات ملتقى الجهود المبذولة من الباحثين في عدة مجالات من المعرفة، والذي من خلاله يتم تطوير وبناء تقنيات تتعامل مع البيانات وأشكالها المتعددة وأنواعها المختلفة بهدف مواجهة المشكلات في مجالات مختلفة كالهندسة، الأعمال، الصناعة، الطب والعلوم.يجمع تنقيب البيانات بين عدة علوم كالإحصاء وتعليم الآلة وقواعد البيانات وتقنيات الإظهار المرئي، ويتجلى هذا الجمع في مراحل وخطوات تنقيب البيانات بدءاً من تجهيز وتجميع البيانات وحتى النتيجة النهائية والتي تختلف حسب أهداف وأغراض التنقيب.**(5)**

تنقيب البيانات يستخدم بعض تقنيات تعليم الآلة مثل الشبكات العصبية وشجرة القرار، ويختلف هدف تنقيب البيانات عن هدف تعليم الآلة، فتعليم الآلة هدفه إعطاء الحواسيب القدرة على تنفيذ مهام يقوم بها البشر عبر تعليمها، بمعنى آخر: استبدال الدور البشري، ولكن تنقيب البيانات هدفه مساعدة الدور البشري ودعمه وليس استبداله.يتمثل الدور الرئيسي لقواعد البيانات في حفظ البيانات والحصول عليها عند الحاجة، بينما دور تنقيب البيانات يتمثل في القدرة على قراءة هذه البيانات وتحليلها للمساعدة في اتخاذ القرار المناسب. تقنيات الإظهار المرئي يتم استخدامها كأداة في مرحلة تحضير البيانات أو مرحلة ما بعد تنقيب البيانات لإظهار النتائج.**(5)**

استخدام تقنيات تنقيب البيانات في المجالات الصحية:

مقدمة:

سرعان ما أصبحت السجلات الصحية الإلكترونية (EHR) أكثر شيوعًا بين مرافق الرعاية الصحية. مع زيادة إمكانية الوصول إلى كمية كبيرة من بيانات المرضى ، يمكن لمقدمي الرعاية الصحية الآن تحسين كفاءة ونوعية مؤسساتهم باستخدام استخراج البيانات. منذ تسعينيات القرن الماضي ، استخدمت الشركات تعدين البيانات لأشياء مثل سجل الائتمان والكشف عن الاحتيال. والآن ، بدأ عدد من مؤسسات الرعاية الصحية بمشاهدة الفوائد المحتملة لتنقيب البيانات والتحليلات التنبؤية.

في مجال الرعاية الصحية، أثبتت عملية استخراج البيانات فعاليتها في مجالات مثل الطب التنبئي ، وإدارة علاقات العملاء، واكتشاف الاحتيال وإساءة الاستخدام، وإدارة الرعاية الصحية وقياس فعالية بعض العلاجات. الغرض من استخراج البيانات، سواء كان يتم استخدامه في الرعاية الصحية أو الأعمال التجارية، هو تحديد أنماط مفيدة ومفهومة من خلال تحليل مجموعات كبيرة من البيانات. وتساعد أنماط البيانات هذه على التنبؤ باتجاهات الصناعة أو المعلومات ، ثم تحديد ما يجب فعله حيالها.

الاشكال الاحصائية لتوزيع البيانات:

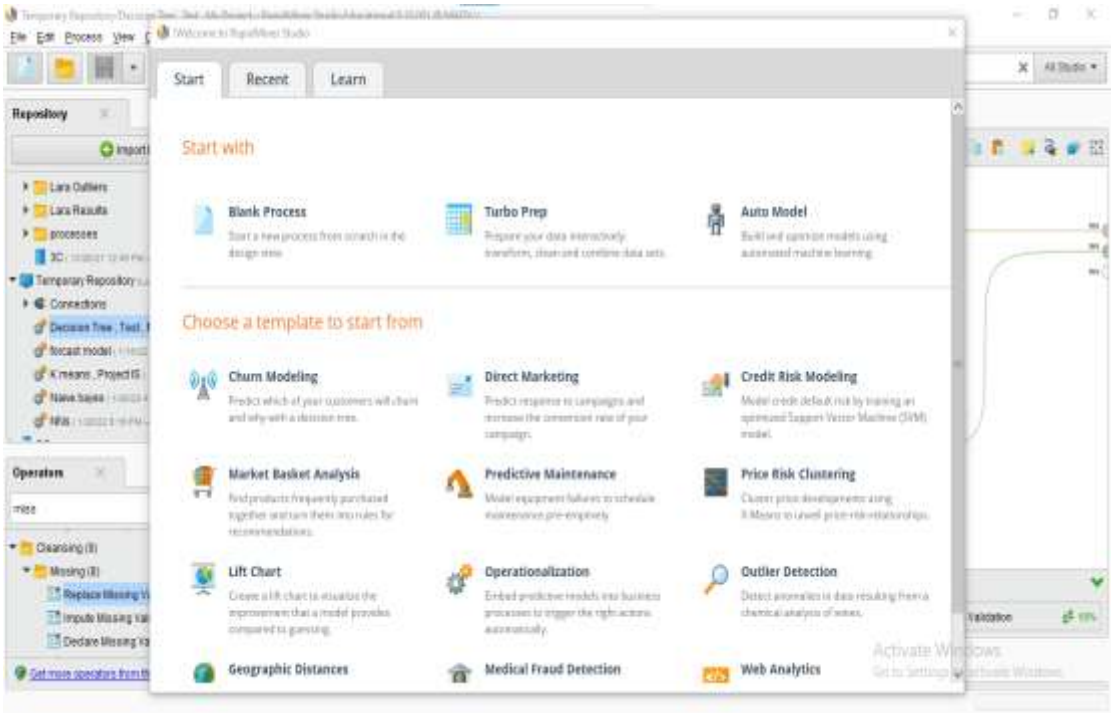
مقدمة عن برنامج (Rapid Miner):

يجمع Rapid Miner Studio بين التكنولوجيا وقابلية التطبيق لتقديم خدمة سهلة الاستخدام دمج أحدث تقنيات التنقيب عن البيانات وكذلك الراسخة.تعريف تتم عمليات التحليل باستخدام Rapid Miner Studio عن طريق سحب وإفلات المشغلين ، تحديد المعلمات والجمع بين العوامل .صفحة البداية :بمجرد دراسة البرامج التعليمية ، يمكنك تحديد خطوتك التالية بمساعدة الصفحة الرئيسية :إذا كنت تريد إرشادات إضافية ، أو ترغب في تسريع عملية إعداد البيانات وبناء النماذج ، فجرّب أداة Turbo Prep ، وأداة Rapid Miner لإعداد البيانات التفاعلية ، وحل Auto Model ، Rapid Minerللتعلم الآلي الآلي. إذا كنت ترغب في رؤية المزيد من الأمثلة ، فاختر منأحد القوالب الموجودة في مستودع العينات ، إذا كنت تريد القيام بذلك بنفسك ، فقم بإنشاء عملية (فارغة) جديدة من البداية في Design View ، ويمكنك فتح صفحة البدء في أي وقت عن طريق تحديد ملف عملية جديدة[1]..

1. الشكل يوضح شعار برنامج شكل شعار Rapid miner

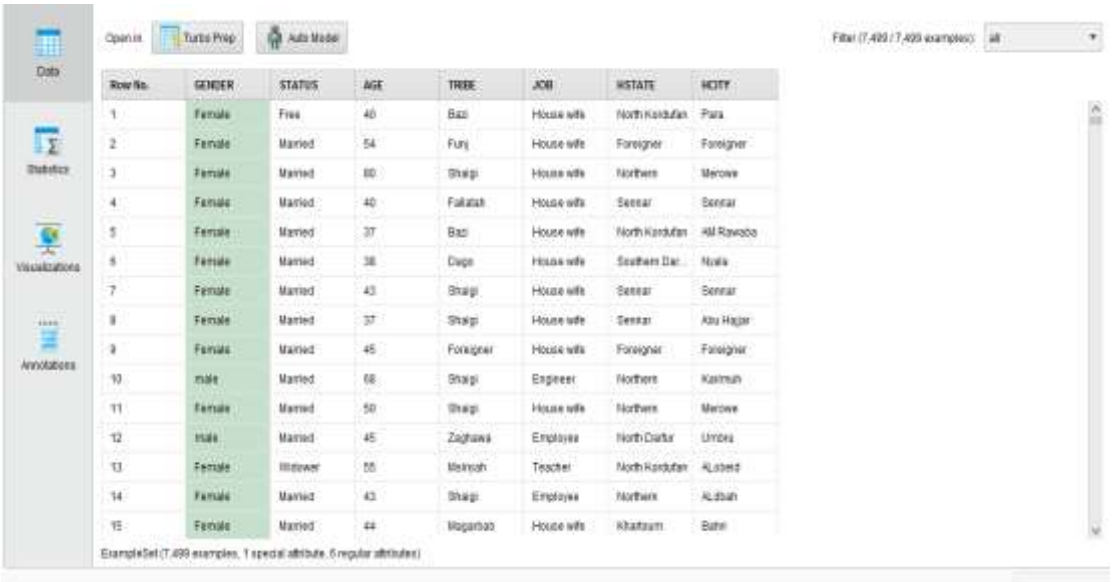
شكل رقم (1) يوضح شعارRapid min

2. الشكل يوضح مجموعة من القوالب الموجودة في مستودع العينات ، إذا كنت تريد القيام بذلك بنفسك ، فقم بإنشاء عملية (فارغة) جديدة من البداية في Design View ، ويمكنك فتح صفحة البدء في أي وقت عن طريق تحديد ملف عملية جديدة.



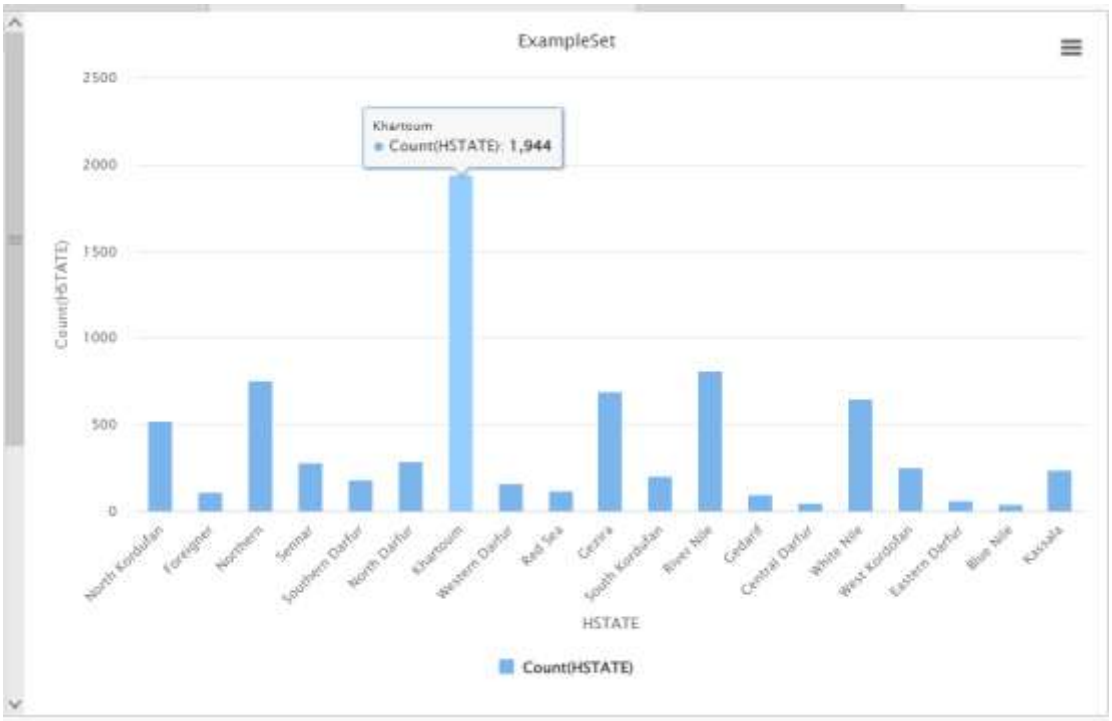
شكل (2) صفحة البداية لـRapid Miner

3. الشكل يوضح عينة الدراسة داخل برامج Rapid Miner حيث تم تحديد النوع (labal)



الشكل (3) يوضح عينة الدراسة داخل برامج Rapid Miner . **التاريخ 2021/9/15م-المصدر مستشفي الذرة إدارة الاحصاء**(6)

4. الشكل يوضح عدد الاصابة في ولايات السودان حيث ان ولاية الخرطوم الاكثر إصابة بعدد 1944 اصابة بين ان ولاية النيل الازرق الاقل إصابة بعدد 45 اصابة .

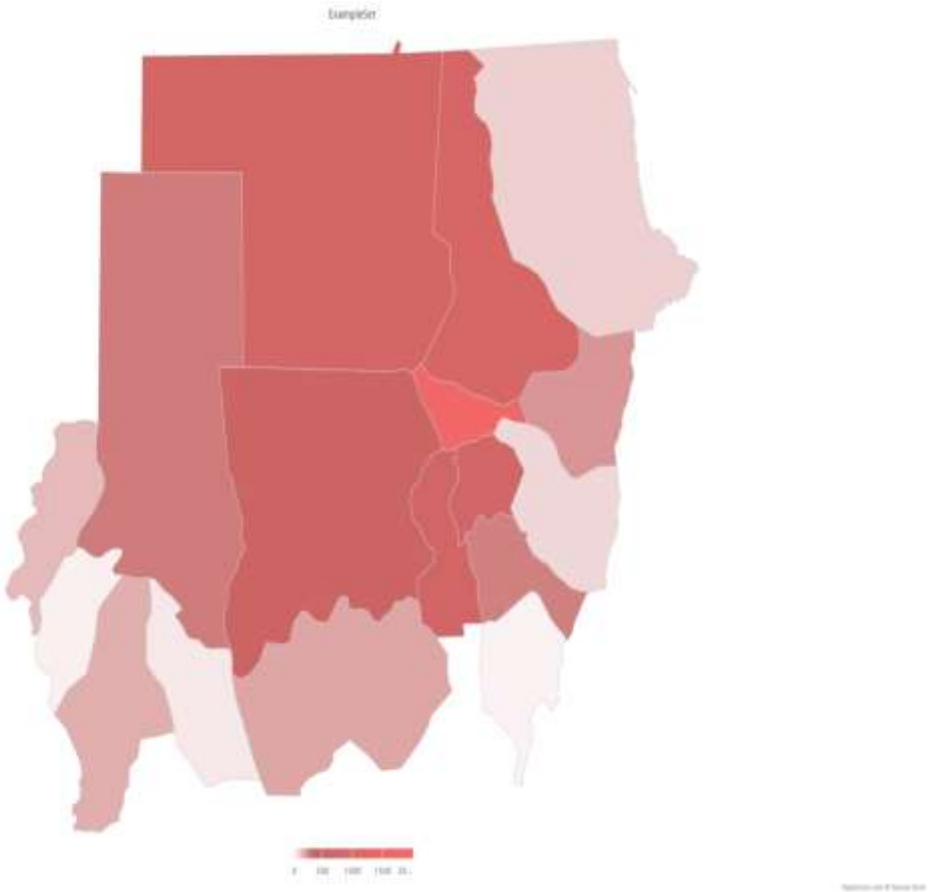


الشكل رقم (3) يوضح الولايات السودانالتاريخ 2021/9/15م-المصدر مستشفي الذرة إدارة الاحصاء(6)

5. الشكل يوضح إحصائيات حسب الولايات

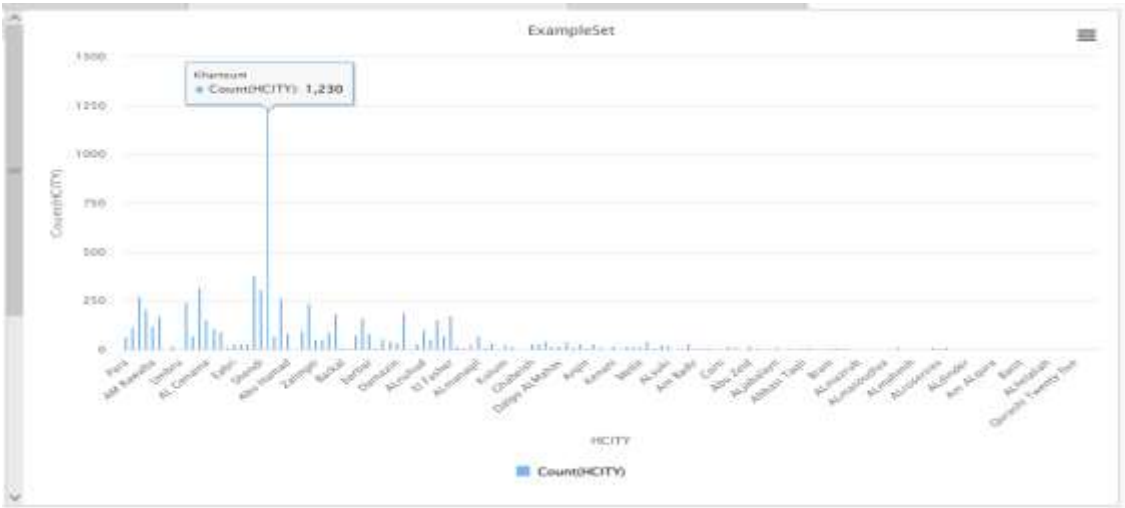
الشكل رقم (5) إحصائيات الولايات, **المصدر مستشفي الذرة الخرطوم إدارة الاحصاء**.(6)

6. شكل يوضح خريطة انتشار الحالات في الولايات:



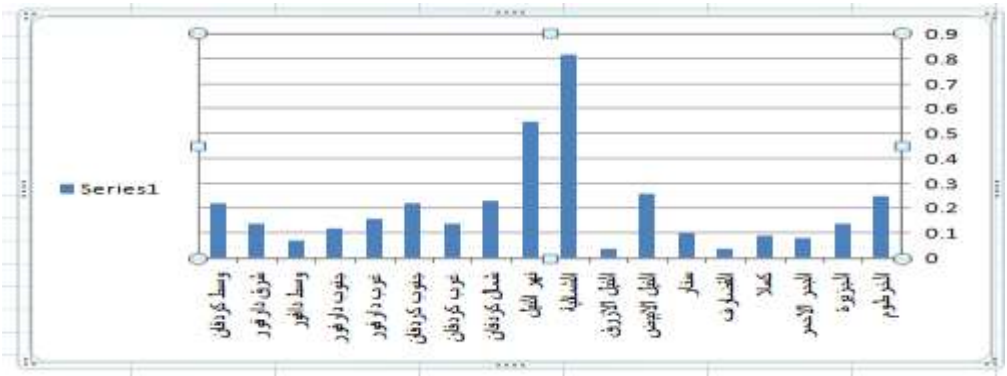
شكل رقم (6) خريطة انتشار الحالات في الولايات, **المصدر مستشفي الذرة الخرطوم إدارة الاحصاء**. (6)

7. الشكل يوضح عدد الاصابة حسب المحليات بالنسبة للولايات وكان العدد الاكبر في محلية الخرطوم بعدد 1230إصابة .



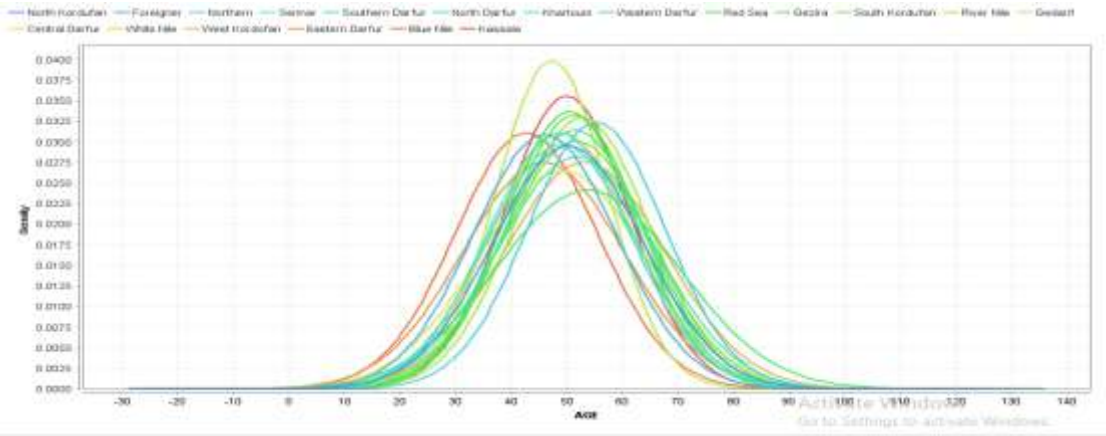
الشكل رقم (7) يوضح الاصابة حسب المحليات, **المصدر مستشفي الذرة الخرطوم إدارة الاحصاء**. (6)

8. الشكل يوضح الخريطة تظهر انتشار الحالات في ولايات السودان بدرجات ألوان مختلفة ، فنجد أن الخرطوم كان عدد سكانها = 7,687,500 نسمة وعدد المصابين =1944, و ولايه الجزيرة عدد السكان يساوي =4,926,600 و عدد المصابين =695 و ولاية البحر الاحمر عدد السكان =1,447,800 وعدد المصابين =118و ولاية كسلا عدد السكان =2,438,800 وعدد المصابين =238 و ولاية القضارف عدد السكان =2,108,500 و عدد المصابين =101 و ولايه سنار عدد السكان = 1,847,500 و عدد المصابين = 282و ولاية النيل الابيض عدد السكان =2,410,300 و عدد المصابين = 647 و ولاية النيل الازرق عدد السكان =1,080,700 و عدد المصابين = 45 و ولاية الشمالية عدد السكان =913,500 و عدد المصابين =757 و ولايه نهر النيل عدد السكان = 1,472,300 و عدد المصابين = 814و ولاية شمال كردفان عدد السكان = 2,206,800 و عدد المصابين = 524 و ولاية غرب كردفان عدد السكان = 1,737,700 و عدد المصابين = 257و ولاية جنوب كردفان عدد السكان =1,263,400 و عدد المصابين =205 و ولاية وسط كردفان عدد السكان =2,296,100 و عدد المصابين = 524 و ولاية غرب دارفور و عدد السكان =995,200 و عدد المصابين = 164 و ولاية جنوب درافور و عدد السكان = 3,672,400 و عدد المصابين = 186 و ولاية شرق درافور عدد السكان = 1,547,800 وعدد المصابين = 65 و ولاية وسط دافور و عدد السكان =729,900 وعدد المصابين = 52 , و عدد السكان الكلي = 40,782,700 نسمة حسب آخر إحصائية بتاريخ 1 يوليو 2017 و عدد المصابين 7499.



الشكل رقم (8) توزيع عدد المصابين علي عدد السكان في كل ولاية.

9. خوارزمية (Naïve Bayes) رسم بياني يوضح العلاقة بين الأعمار وتوزيعها في الولايات فيNaïve Bayes , حيث كانت اكثر الولايات توزرعا ولاية الخرطوم .



الشكل رقم (9)يوضح خوارزمية (Naïve Bayes) , **المصدر مستشفي الذرة الخرطوم إدارة الاحصاء.** (6)

10. الشكل وضح دقة (Naïve Bayes :)لخوارزميةالعدد النسبي للأمثلة المصنفة بشكل صحيح .

accuracy: 99.16%

	true Nort...	true Fore...	true Nort...	true Sen...	true Sout...	true Nort...	true Khar...	true Wes
pred. No...	155	0	0	0	0	0	0	0
pred. For...	0	34	0	0	0	0	0	0
pred. No...	0	0	226	0	0	0	0	0
pred. Se...	0	0	0	85	0	0	0	0
pred. So...	0	0	0	0	56	0	0	0
pred. No...	2	0	0	0	0	88	0	0
pred. Kh...	0	0	1	0	0	0	583	0
pred. We...	0	0	0	0	0	0	0	49
pred. Re...	0	0	0	0	0	0	0	0
pred. Ge...	0	0	0	0	0	0	0	0
pred. So...	0	0	0	0	0	0	0	0

الشكل رقم (10)يوضح خوارزمية (Naïve Bayes) , **المصدر مستشفي الذرة الخرطوم إدارة الاحصاء**.(6)

الخاتمة :

بعد الدراسة في بيانات سلطان الثدي في مستشفي الذرة الخرطوم فقد تبين انولاية الخرطوم الاكثر إصابة بعدد 1944 اصابة بين ان ولاية النيل الازرق الاقل إصابة بعدد 45 اصابةوان الولاة الشمالية هي الاكثر إصابة مقارنتا بعدد عدد الاصابات نسبتا لي عدد السكان, وقد وصينا انه يجب تخزين بيانات المريض بطريقة أفضل لتكون في متناول اليد ، مما يساعد في استكمال عملية البحث في هذا المجال والتوعية والإرشاد للكشف المبكر عن هذا المرض و تحليل بيانات الاصابة في الاعوام القادمة و مقارنتها مع الاعوام السابقة.

النتائج :

بعد إستخدام الطرق الإحصائية في برنامج (Rapid Miner) تم التوصل للنتائج الاتية

1. ان ولاية الخرطوم الاكثر إصابة من حيث عدد الاصابات بعدد 1944 اصابة بين ان ولاية النيل الازرق الاقل إصابة من حيث عدد الاصابات بعدد 45 اصابة.
2. ان الولاية الشمالية هي الاكثر اصابة مقارنتا بعدد السكان مع نسبة الاصابة.
3. ان خوارزمية (Naïve Bayes) توضح العلاقة بين الأعمار وتوزيعها في الولايات فيNaïve Bayes , حيث كانت اكثر الولايات توزعا ولاية الخرطوم.
4. اندقة (Naïve Bayes :)لخوارزميةالعدد النسبي للأمثلة المصنفة بشكل صحيح أوبعبارة أخرى النسبة المئوية للتنبؤات الصحيحة حيث كانت الدقة 99.16%.
5. ان عدد الاصابات حسب المحليات كان العدد الاكبر في محلية الخرطوم بعدد 1230إصابة .

التوصيات:-

1. إستخدم برنامج آخر غير (Rapid miner) وقارن بينها وبين برنامج النتائج مع بعضها البعض .
2. يجب تخزين بيانات المريض بطريقة أفضل لتكون في متناول اليد ، مما يساعد في استكمال البحث في هذا المجال .
3. التوعية والإرشاد للكشف المبكر عن هذا المرض
4. تحليل بيانات الاصابة في الاعوام القادمة و مقارنتها مع الاعوام السابقة.

المصادرو المراجع :

زياد عبدالكريم القاضي (2004)."مقدمة في تصميم قواعد البيانات" ، "دار صفاء للطباعة والنشر والتوزيع" ،2004م.

ياسر مطيع، محمد الراميني، تامر جلال، محمد نصرالله (2005). "أساسيات قواعد البيانات" ، "دار صفاء للطباعة والنشر والتوزيع" ،2005م.

مراد شلباية،نهلة درويش،وائل أبو مغلي (2005). "مفاهيم أساسية في قواعد البيانات" ، "دار المسيرة للنشروالتوزيع و الطباعة" ،2002م.

حجازي، محمد عثمان (2006). " محاضرات في برمجة قواعد البيانات" , جامعة حائل كلية التربية، "2006

أروى عيسى الياسري (2006). "إستخراج البيانات Data Miningاتجاه جديد في استرجاع المعلومات" , "مجلة المعلوماتية – العدد 16"

مستشفي الذرة إدارة الاحصاء. التاريخ 15/9/ 2021م.

Liao M-N, Chen S-C, Lin Y-C, Chen M-F, Wang C-H, Jane S-W (2025). Education and psychological support meet the supportive care needs of Taiwanese women three months after surgery for newly diagnosed breast cancer: A non-randomised quasi-experimental study. International journal of nursing studies. 2025; 51(3): 390-9.

Nikbakhsh N, Moudi S, Abbasian S, Khafri S. Prevalence of depression and anxiety among cancer patients. Caspian J Intern Med. 2025; 5(3): 167-70.

Haghighi M, Rahmati-Najarkolaei F, Ansarian A. Correlation between Spiritual Wellbeing and Religious Orientation among Staffs of one Military Medical University. Journal of Health Policy and Sustainable Health. 2018; 1(4): 137-40.

http://hdl.handle.net/123456789/2076

http://repository.sustech.edu/handle/123456789/2076

https://mafhome.com/%D9%85%D8%A7%D9%87%D9%8A%D8%A7%D9%84%D8%A8%D9%8A%D8%A7%D9%86%D8%A7%D8%AA%D8%A7%D9%84%D8%B6%D8%AE%D9%85%D8%A9big-data%D8%9F