تصميم إطار عمل باستخدام أدوات تنقيب البيانات للتنبؤ بمرض السكري

سلمي عثمان محمد قسم الله

أستاذ مساعد بكلية علوم الحاسوب وتقانة المعلومات- جامعة السودان المفتوحة

المستخلص:

تكمن مشكلة الدراسة الرئيسية في إيجاد صعوبة لإتخاذ حلول للتحكم في مرض السكري مما ينتج مشاكل فرعية تتمثل في إنتشار مرض السكري بصورة كبيرة في الآونة الأخيرة وعدم التحكم في مرحلة المرض على حسب ظروف وأحوال المناطق المعينة والأزمنة المحددة. يتمثل الهدف الرئيسي للدراسة في تصميم وتطبيق عمل معين بالتنبؤ بمرض السكري باستخدام أدوات تنقيب البيانات.و هنالك أهداف فرعية منها التحقق من إمكانية استخدام أدوات التنقيب عن البيانات للتنبؤ بمرض السكري.و تحديد نوعية المرض وفي اي مرحلة (يساعد التنقيب في التنبؤ بالمرض و تحديد مرحلة هذا المرض) ووضع الحلول المناسبة وإتخاذ سبل الوقاية اللازمة للحد من إنتشار المرض. تأتي أهمية الدراسة من أهمية موضوع مرض السكري وذلك بأنه من أكثر الأمراض انتشاراً، وقد تكون أمراض وراثية وهي من الأمراض الدائمة للشخص أي صعب العلاج منه نهائياً وهي ملازمة المريض ولها تأثيرات جانبية لا بد من التركيز عليها(العيون ,الأسنان , وفي بعض الأحيان العظام ومشاكل الكلي وبتر الأعضاء). يتبع هذا النظام المنهج الوصفي التحليلي والتطبيقي وتعمل منهجية البحث على تحقيق أغراض الدراسة عن طريق أساليب متنوعة منها: النهج الكمي والنوعي. توصلت نتائج الدراسة للإكتشاف المبكرللمرض مما يقلل الأثار الجانبية وقلة التكلفة وتحسين وضع التنبؤ باستخدام تنقيب البيانات. توصي الدراسة بالحاجة لتوفر سجلات بينات المرضي وذلك بجمع بيانات حقيقية من المؤسسات الصحية بالسودان واستخدام نماذج أخري عدا الإنحدار الخطي وشجرة القرار.

Abstract:

The main problem of the study lies in finding it difficult to take solutions to control diabetes, which results in sub-problems represented in the large spread of diabetes in recent times and the lack of control in the stage of the disease according to the conditions and conditions of specific regions and specific times. The main objective of the study is to design and implement a specific work to predict diabetes using data mining tools. this disease) and develop appropriate solutions and take the necessary preventive measures to reduce the spread of the disease. The importance of the study comes from the importance of the topic of diabetes, as it is one of the most prevalent diseases, and it may be hereditary diseases, and it is one of the permanent diseases of the person, which is difficult to treat permanently. Kidney and amputation) This system follows the descriptive analytical and applied approach and the research methodology works to achieve the objectives of the study through a variety of methods, including: quantitative and qualitative approaches. The results of the study reached the early detection of the disease, which reduces side effects, lower costs, and improves the prediction situation using data mining. The study recommends the need for the availability of patient data records by collecting real data from health institutions in Sudan and using models other than linear regression and decision tree.

مقدمة:

تعد التكنولوجيا الحديثة من أكثر الموضوعات التي أثرت في بيئة العمل بصورة واضحة، فقد سمحت بدخول قدرات وإمكانيات جديدة لدعم كافة النشاطات الحديثة، إذ أصبحت هذه التكنولوجيا عاملا مهما في تغيير ثقافة المنظمات والشركات إلى ثقافة معتمدة على التكنولوجيا سواء في إدارتها أو إستعمالاتها أو في طرق اتخاذ القرار ومن هذه الأدوات التنقيب عن البيانات.

التنقيب هو عمليه بحث محوسب ويدور عن معرفة من البيانات دون فرضيات مسبقة عما يمكن. ويعرف التنقيب في البيانات على أنه عملية تحليل كمية بيانات، لإيجاد علاقة منطقية تلخص البيانات بطريقة جديدة تكون مفهومة ومفيدة.

ظهر التنقيب في البيانات في اواخر الثمانيات واثبت وجوده كأحد الحلول لتحليل كميات ضخمه من البيانات وذلك بتحويلها من مجرد معلومات متراكمة وغير مفهومة إلى معلومات قيمة يمكن استغلالها و الإستفادة منها، كما يعرف أنه تحليل كمية من البيانات عادة ماتكون كبيرة لإيجاد علاقة منطقية تلخص البيانات بطريقة جديدة تكون مفهومة ومفيدة.

مشكلة الدراسة

تكمن مشكلة البحث الرئيسية في إيجادصعوبة لإتخاذ حلول للتحكم في مرض السكري وهنالك مشاكل فرعية تتلخص في:

1. إنتشار مرض السكري بصورة كبيرة في الآونة الأخيرة.
2. عدم التحكم في مرحلة المرض على حسب المناطق المحددة أو ظروف وأحوال الأزمنة المعينة.

أهداف الدراسة:

يتمثل الهدف الرئيسي للبحث في تصميم وتطبيق عمل معين بالتنبؤ بمرض السكري باستخدام أدوات تنقيب البيانات مما ينتج أهداف فرعية تتمثل في:

1. التحقق من إمكانية استخدام أدوات التنقيب عن البيانات للتنبؤ بمرض السكري.
2. تحديد نوعية المرض وفي أي مرحلة (يساعد التنقيب في التنبؤ بالمرض في بعينها وتحديد مرحلة هذا المرض).
3. وضع الحلول المناسبة واتخاذ سبل الوقاية اللازمة للحد من انتشار المرض

أهمية الدراسة:

تأتي أهمية الدراسة من أهمية موضوع مرض السكري وذلك بالآتي:

1. مرض السكري من أكثر الأمراض انتشاراً وقد تكون أمراض وراثية.

2. من الأمراض الدائمة للشخص أي صعب العلاج منه نهائياً وهي ملازمة المريض ولها تأثيرات جانبية لا بد من التركيز عليها(العيون ,الأسنان,وفي بعض الأحيان العظام ومشاكل الكلى وبتر الأعضاء).
3. الإستفادة من البيانات الموجودة لدى المؤسسات الصحية في التنبؤ بنوع مرض السكري النوع (الأول، الثاني).
4. تقليل التكلفة المادية والزمنية في التنبؤ بتحديد نوع مرض السكري.

الإطار النظري:

ينتج مرض السكري عن فقدان هرمون الأنسولين الذي تفرزه خلايا خاصة (خلايا – ب) في البنكرياس أو عن قلة كمية هذا الهرمون أو قلة إستجابة خلايا الجسم له في كثير من الحالات.

وهرمون الأنسولين له فاعلية أساسية في عمليات الاستقلاب والتعامل مع الغذاء بشكل عام ومع السكر بشكل خاص لإنتاج الطاقة اللازمة للجسم ولبناء الأنسجة المختلفة، ويؤدي فقدانه الكمي أو النوعي إلى تراكم السكر في الدم بدرجات لم تتعود عليها أنسجة الجسم مما يتسبب في إحداث اختلالات عديدة قد تظهر على المدى القريب أو البعيد.

ويندرج تحت ما يسمى بمرض السكري عدة أنواع تختلف عن بعضها البعض اختلافاً كثيراً في الأسباب وطرق العلاج، ونورد فيما يلي أنواع هذا المرض كما هو متفق عليه من تسميات وتصنيفات لدى المؤسسات الطبية العالمية المتخصصة في مرض السكري. (Cooper and Schindler 1995, Agrawal and Srikant) (2003)

أنواع مرض السكري

1. السكري من النوع الأول (Diabetes Mellitus Type 1)
2. السكري من النوع االثاني (Diabetes Mellitus Type 2)
3. سكري الحمل (Gestational Diabetes)
4. أنواع أخرى:

أ. السكري الناتج عن بعض أمراض البنكرياس.

ب. السكري الناتج عن إختلالات هرمونية وخصوصاً في الغدد النخامية والكظرية وخلايا (1) في البنكرياس.

ت. السكري الناتج عن بعض الأدوية.

ث. أنواع أخرى نادرة.

إن المرضى المصابين بداء السكري يصنفهم الأطباء المتخصصون إلى الأقسام الأربعة التالية:

1. المرضى ذوو الاحتمالات الكبيرة جداً للمضاعفات الخطيرة بصورة مؤكدة طبياً وتتميز أوضاعهم المرضية بحالة أو أكثر مما يأتي:

أ. حدوث هبوط السكر الشديد خلال الأشهر الثلاثة التي سبقت شهر رمضان.

ب. المرضى الذين يتكرر لديهم هبوط وارتفاع السكر بالدم.

ت. المرضى المصابون بحالة (فقدان الإحساس بهبوط السكر)، وهي حالة تصيب بعض مرضى السكري، وخصوصاً من النوع الأول، الذين تتكرر لديهم حالات هبوط السكر الشديد ولفترات طويلة.

ث. المرضى المعروفون بصعوبة السيطرة على السكري لفترات طويلة.

ج. حدوث مضاعفة (الحماض السكري الكيتوني) أو مضاعفة (الغيبوبة السكرية الأسمولية) خلال الشهور الثلاثة التي سبقت شهر رمضان.

ح. السكري من النوع الأول.

خ. الأمراض الحادة الأخرى المرافقة للسكري.

د. مرضى السكري الذين يمارسون مضطرين لأعمال بدنية عنيفة.

ذ. مرضى السكري الذين يجرى لهم غسيل كلى.

ر. المرأة المصابة بالسكري أثناء الحمل. (Cooper and Schindler,2003)

2. المرضى ذوو الاحتمالات الكبيرة نسبياً للمضاعفات نتيجة الصيام والتي يغلب على ظن الأطباء وقوعها وتتميز أوضاعهم المرضية بحالة أو أكثر مما يأتي:

أ. الذين يعانون من إرتفاع السكر في الدم كأن يكون المعدل 180 – 300مغم/ دسل، (10ملم – 16 .5 ملم) ونسبة الهيموغلوبين المتراكم (المتسكر) التي تتجاوز 10 %.

ب. المصابون بقصور كلوي.

ت. المصابون بإعتلال الشرايين الكبير (كأمراض القلب والشرايين).

ث. الذين يسكنون بمفردهم والذين يعالجون بواسطة حقن الأنسولين أو العقارات الخافضة.

ج. الذين يعانون من أمراض أخرى تضيف أخطاراً إضافية عليهم.

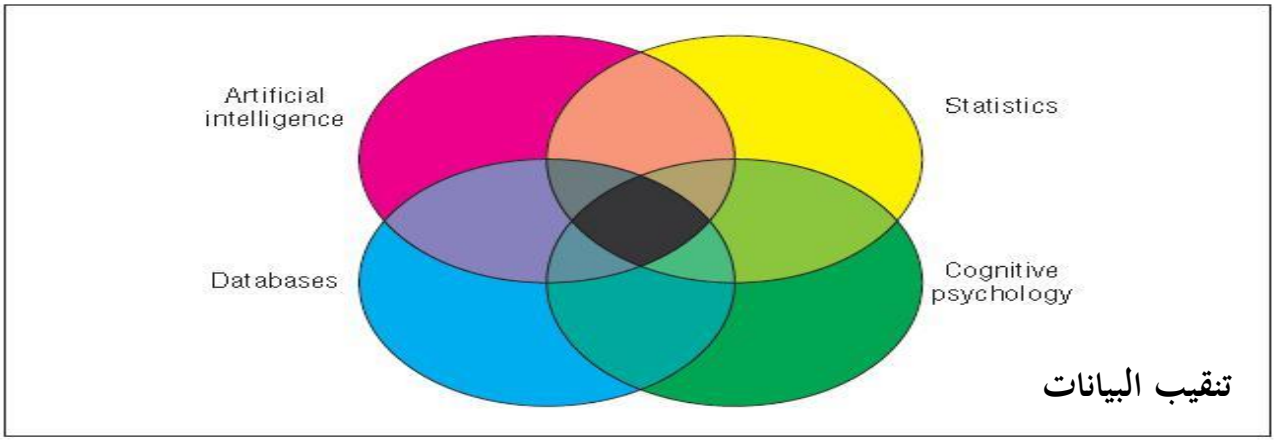
ح. كبار السن المصابون بأمراض أخرى مثل السرطان.

خ. المرضى الذين يتلقون علاجات تؤثر على العقل.

3. المرضى ذوو الاحتمالات المتوسطة للتمرض للمضاعفات نتيجة الصيام ويشمل ذلك مرضى السكري ذوي الحالات المستقرة والمسيطر عليها بالعلاجات المناسبة الخافضة للسكر التي تحفز خلايا البنكرياس المنتجة للأنسولين.
4. المرضى ذوو الاحتمالات المنخفضة للتعرض للمضاعفات نتيجة الصيام ويشمل ذلك مرضى السكري ذوي الحالات المستقرة والمسيطر عليها بمجرد الحمية أو بتناول العلاجات الخالصة للسكر التي لا تحفز خلايا البنكرياس للأنسولين بل تزيد فاعلية الأنسولين الموجود لديهم.

الجمع بين علم النفس المعرفي والذكاء الاصطناعي وقواعد البيانات والإحصاءات

شكل رقم (1) الجمع بين علم النفس المعرفي والذكاء الاصطناعي وقواعد البيانات والإحصاءات



تتوفر الآن كميات هائلة من سجلات البيانات في مجالات العلوم والأعمال والصناعة والعديد من المجالات الأخرى". للتعرف على هذه البيانات وتحليلها وتوظيفها في النهاية، تم إقتراح تقنية متعددة التخصصات تسمى تنقيب البيانات. (Fu, Y., 1997–18 ,4 :16-)، وهي إكتشاف المعرفة والمعلومات من البيانات.(Kantardzic, M., 2003,1/3/2003)

يعد تنقيب البيانات عملية تحديد الميزات أو العلاقات أو الأنماط أو النماذج المثيرة للاهتمام من قواعد البيانات الكبيرة من أجل إدارة أعمالك بشكل أفضل.

أ. كما هو موضح في الشكل 1، يعد التنقيب عن البيانات هو الجزء الأساسي والجزء الرئيسي من إكتشاف المعرفة في قاعدة البيانات. يحتوي إكتشاف المعرفة عادة على الخطوات التالية: جمع البيانات، تصحيح البيانات، تحويل البيانات، إستكشاف الأنماط (تنقيب البيانات)، تفسير النتائج، التقييم واستخدام المعرفة المكتشفة. Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P., (Magazine, 17:3, 186-193).

ب. "العديد من البائعين والمستشارين والمحللين يجعل تنقيب البيانات يبدو معقدًا وصعبًا وبائعًا ومكلفًا. قد يكون الأمر في بعض الأحيان معقدًا (يتضمن العديد من الأجزاء)، لكن لا يلزم أن يكون غامضاً أو صعباً.

ت. تنقيب البيانات يعني ببساطة: العثور على أنماط في البيانات الخاصة بك والتي يمكنك استخدامها لإجراء أعمالك بشكل أفضل. . Kantardzic, M. (2003) Data Mining: Concepts, Models, Methods, and Algorithms. John Wiley & Sons, Hoboken

تنقيب البيانات هو أكثر من مجرد تحليل البيانات التقليدية". يستخدم وسائل التحليل التقليدية وتلك المرتبطة بالذكاء الاصطناعي. تنقيب البيانات هو نظرة أو نهج فريد من نوعه لتحليل البيانات. الهدف هو تقديم أسئلة أكثر للحصول على دقة الإجابات. يمكن التحقق من صحة المجاملات التي تم تحقيقها من خلال تنقيب البيانات من خلال التحليل التقليدي.
(B.J., 1999, 1/7/1999)

هنالك فرق بين إكتشاف المعرفة وتنقيب البيانات

تنقيب البيانات وهي تقنيات للحصـــول على بعض سمات الكيان من مجموعة من البيانات على مكون فردي إلى كل من الخصائص / الميزات التي تم الحصول عليها من قبل المراقبة. .

(Ohsuga S., (2005), 25-27 July, IEEE, 1, 7- 12)

يشـير مصـطلح "إكتشـاف المعرفة في قواعد البيانات" إلى العملية الكاملة لإكتشـاف المعرفة القيمة من البيانات. إن إكتشـــاف المعرفة في قواعد البيانات هو عملية التعرف على الأنماط / النماذج المناســـبة والرواية والتي يمكن أن تكون عملية، وفي النهاية مفهومة في البيانات. (Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996) From Data Mining to Knowledge Discovery in Databases. American Association for Artificial Intelligence, 17, 36-51

يعد (تنقيب) البيانات العنصر الأساسي في الإجراء الأكثر شيوعًا لإكتشاف المعرفة في قواعد البيانات. ((Read, B.J.,2000,11/12/2000

حقل إكتشـــاف المعرفة وتنقيب البيانات يعتمد على النتائج المســتخلصــة من الإحصــاءات وقواعد البيانات والذكاء الاصـطناعي لبناء الأدوات التي تتيح للمسـتخدمين إكتسـاب نظرة ثاقبة من مجموعات البيانات الضخمة". (Fu, Y., (1997),20–18 ,4 :16)

في الواقع مصطلح إكتشاف المعرفة هو إجراء تكراري وتفاعلي يتضمن مختلف المراحل وتحديد المشـــكلة وفهم مجال التطبيق، وهي الخطوة الأســـاســـية لفهم مجال التطبيق. من الواضـــح أن هذه الخطوة هي شـرط مسـبق لاسـتخراج المعرفة القيمة واختيار تقنيات تنقيب البيانات المناسـبة وفقاً لهدف التطبيق وطبيعة البيانات. ((Mitra, S., and Acharya, T., (2014) 02/13/2014 13:58

التنقيب على مجموعة البيانات المخفضـــة يجب أن يكون أكثر كفاءة، لكنه يخلق النتائج التحليلية نفســـها أو تقريباً يمكن لتطبيق التقنيات تقليل عدد القيم لميزة مســـتمرة معينة عن طريق فصـل سـلسـلة الخاصـية إلى فواصـل زمنية. وهذه العملية تعرف بتقديرية البيانات:،و يمكن تطبيق علامات الفاصـل الزمني لاسـتبدال قيم البيانات الحقيقية . Han, J. and Kamber, M. (2006) Data (Mining:, 5, 1-18.)

إكتشاف المعرفة يؤدي إلى تنوير الإتجاهات أو الحقائق الموجودة أو التاريخية، والتنبؤ بالمستقبل، ومساعدة صناع القرار على وضع إستراتيجية من الحقائق والمعلومات المستخرجة. Han, J. and (Kamber, M. (2006) Data Mining:, 5, 1-18.)

يهدف التعليم الآلي إلى تقديم مستويات متزايدة من الميكنة في إجراء هندسة المعرفة، والاستعاضة عن النشاط البشري الذي يستغرق وقتاً طويلًا بطرق تلقائية تعمل على تحسين الدقة أو الكفاءة من خلال إكتشاف وتطبيق عناصر انتظامية في بيانات التدريب. Langley, P., Simon, H.,1995, 11, 54 – 64)

تنقيب البيانات والإحصاء:

تنوي الإحصائيات وتنقيب البيانات كلاهما إكتشاف بنية البيانات نظراً لتداخل كميات كبيرة من خططهم، متمثلة في الخوارزميات الإحصائية الأساسية وهي التقنيات الوصفية والتصور مثل تحليل الكتلة، تحليل الإرتباط، تحليل التمييز، تحليل العوامل، تحليل الإنحدار، الإنحدار اللوجستي. (Hand, D. J., 1999, 1: 1, 16 – 19)

تقنيات التصنيف الرئيسية هي.

1. شجرة القرار (عرفي).
2. الشبكات العصبية (عرفي).
3. تحليل الإرتباط (عرفي).
4. أقرب تقنيات الجار (عرفي).
5. دعم ناقلات الآلات. (Kantardzic, M., 2003)

شجرة القرار عبارة عن تمثيل بياني لعملية القرار وتتكون هذه الشجرة من العناصر التالية: نقاط القرار، البدائل، نقاط الفرص أو الحدث، حالات الطبيعة، والعوائد. (نبيل محمد مرسي، 2006م، 2011/1/1)

تعتبر شجرة القرارات من الأدوات التي يعتمد عليها متخذ القرار في حل المشكلات والتي تمثل في النهاية القرار الذي سوف نتوصل له لحل المشكلة. (عبد الملك إسماعيل حجر، 2001/1,2011)

ويجب أن تشتمل بيانات شجرة القرار على الاحتمالات الخاصة بالفروع التي تخرج من منابت الأحداث والإيرادات الخاصة بالبدائل المختلفة للمشكلة.)حمدي طه، 1/1/2011,2011)

التنبؤ باستخدام الإنحدار الخطي

الإنحدار أو يسمى التنبؤ Prediction وهو تقدير القيمة المستقبلية لمتغير واحد بناءاً على معرفة قيم متغير أو أكثر، وهنالك عدة أنواع من معامل الإنحدار: http://un.uobasrah.edu.iq/lectures/1988.pdf

1. الإنحدار الخطي Linear Regression تشير تسمية هذا المعامل إلى أنه يتضمن متغير تابع y يعتمد على متغير واحد مستقل x وكلمة خطي تشير إلى أن العلاقة ين المتغيرين y و x هي علاقة خطية.
2. الإنحدار المتعدد Multiple Linear Regression هذا النوع من الإنحدار يتضمن اعتماد المتغير y على أكثر من متغير مستقل مثل x1 و x2... الخ.
3. الإنحدار غير الخطي Non-Linear Regression إذا كانت العلاقة بين المتغير y والمتغيرات المستقلة غير خطية مثل علاقة أسية أو لوغاريثمية أو تربيعية... الخ. وهنالك أنواع أخرى مثل الإنحدار الهرمي Hierarchical Regression والإنحدار التدريجي Stepwise Regression وغيرها.

الإنحدار الخطي هو أداة إحصائية تستعمل لبيان العلاقة بين متغيرين كميين بحيث يمكن توقع قيمة المتغير التابع Dependent variable (y) المسيطر عليه من المتغير المستقل (x) Independent variable المسيطر عليه. على سبيل المثال، إذا كان الباحث يعرف العلاقة بين النسبة المئوية لتراكم المادة الجافة وإنتاجية الحنطة فانه يمكنه التنبؤ بالإنتاجية عن طريق الإنحدار الخطي بمجرد تحديد مستوى تراكم المادة الجافة، بصورة عامة يستعمل الإنحدار للأغراض الآتية:

1. تعد هذه الطريقة تقنية لنمذجة وتحليل البيانات العددية.

2. استغلال العلاقة بين متغيرين للتنبؤ بقيم أحد المتغيرات من خلال قيم المتغير الأخر.
3. التنبؤ وتقدير واختبار فرضية ونمذجة العلاقات السببية.

منهجية الدراسة:

يمكن إجراء الدراسة التحليلية للتنبؤ أو لمعرفة الدرجة التي ترتبط بها المتغيرات ,ويتم التركيز على تحليل الموقف أو المعضلة لتوضيح العلاقات بين المتغيرات، يمكننا تعيين مهام استخراج البيانات في واحدة من فئتين.

أولاً: تنبؤ بيانات التنقيب وذلك "يتضمن استخدام بعض المتغيرات أو الحقول في قاعدة البيانات للتنبؤ بقيم غير معروفة أو مستقبلية للمتغيرات الأخرى المثيرة للاهتمام".

ثانياً: التنقيب عن البيانات الوصفية الذي يركز على إكتشاف الأنماط التي يمكن تفسيرها من قبل الإنسان وتحديد الخصائص العامة للبيانات في قاعدة البيانات التي قد تساعد المهام الوصفية أيضاً في الأبحاث للتنبؤ بها (مثل هذه الدراسة).. نظراً لأن أداتنا في هذه الدراسة هي التنقيب عن البيانات، وسوف نقوم بتطبيق كل من المهام الوصفية والتنبؤية لاستخراج البيانات، فإن الغرض من هذا الدراسة هو التحليل بشكل أساسي ولكن مع بعض الجوانب الوصفية التي تساعدنا في مرحلة التنبؤ الخاصة بنا.

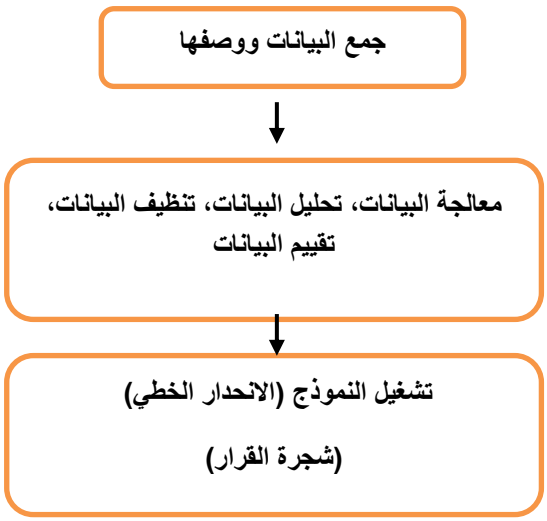
مرحلة تحليل النظام:

يتم جمع البيانات ووصفها وإعدادها وتحليلها، وتم جمع البيانات من مصدر عبر الإنترنت (الرابط). والهدف من هذه الورقة هو دراسة مشكلة التنبؤ بمرض السكري، والتحقق من إمكانية استخدام أدوات التنقيب عن البيانات للتنبؤ بمرض السكري وتحديد الطريقة المناسبة للتنبؤ بالمرض، سيتم تحديد التنبؤ على أساس المقارنة بين الإنحدار الخطي وشجرة القرار.

نموذج الإنحدار الخطي وشـــجرة القرار لـديهم القـدرة على التنبؤ. نســـتخدم أداة Rapid miner،" وهي مصدر مفتوح لبرنامج الترميز ويعطي تحليلات متقدمة وسهلة الاستخدام في عملية استخراج البيانات".

نموذج الإنحدار الخطي

شكل رقم (2) نموذج الإنحدار الخطي

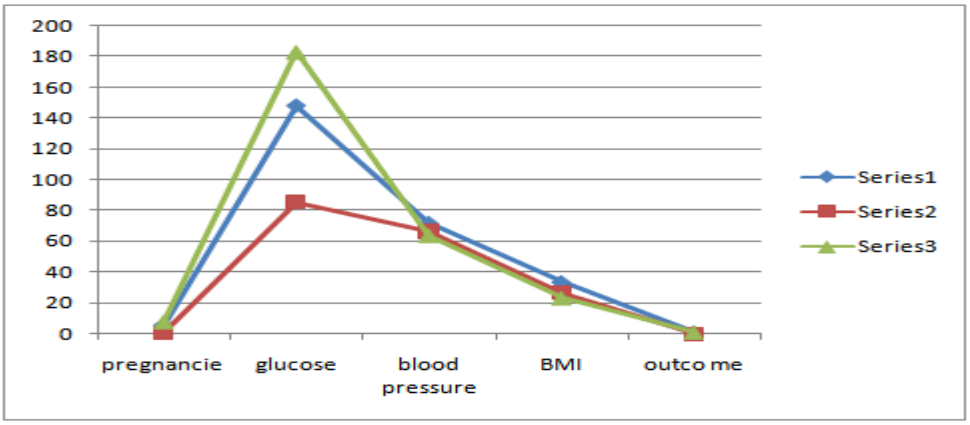


إعداد البيانات:

بعد جمع البيانات اللازمة يجب دمج البيانات وتنظيمها وتحويلها لتكون مناســـبة للتنبؤ بمرض الســكري. لأن قواعد البيانات حســاســة للغاية للبيانات المفقودة وغير المنســقة، هنالك عدد من التقنيات مثل معالجة البيانات الأولية، تنظيف البيانات، تكامل البيانات، تحويل البيانات، والحد من البيانـات، يمكن تطبيق تنظيف البيـانـات لإزالة الشـــواذ، والقيم المفقودة في البيـانـات.تـدمج البيانات من مصـــادر متعددة في مخزن بيانات واضــح، البيانات تتضــمن التحويلات تطبيع/ قياس وبنية المعالم. يتم تحويل البيانات وتوحيدها في هياكل مناسبة للتنقيب.

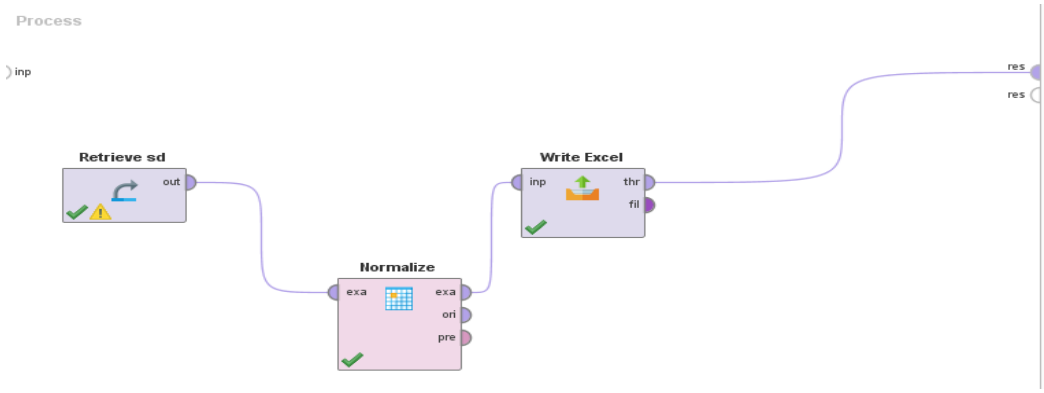
مخطط يوضح الوقت اللازم للتنقيب الحقيقي في الجودة العامة

مخطط رقم (1) الزمن اللازم للتنقيب الحقيقي



يســـــاعد تطبيع بيانات الادخال في تســـــريع مرحلة التدريب ويجب ان تكون جميع البيانات موحدة قبل النمذجة.

شكل رقم (3) معلمة مرشح السمة



أ. نوع مرشح (السمة) (مجموعة فرعية):

تسمح لك هذه المعلمة بتحديد مرشح تحديد السمات، التي لديها عدة خيارات وهي:

أخذ عينات البيانات للتدريب والاختبار:

في هذه المرحلة يجب تقسيم البيانات إلى مجموعات تدريب واختبار. في الواقع يتم تطبيق مجموعة لإنشاء النموذج، ويتم تطبيق مجموعة بيانات الاختبار لاختبار النموذج.

للقيام بهذه المرحلة في Rapid miner يمكنني استخدام عامل تشغيل (تقسيم البيانات) وجعلها مقسمة إلى بيانات 0.5، 0.5 للتدريب

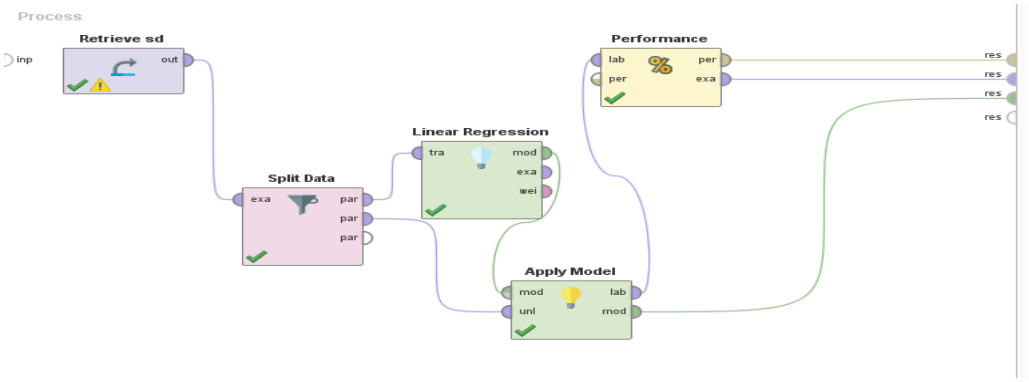
الخطوات:

1. استرداد البيانات(DIABETES).
2. تقسيم البيانات (0.5، 0.5).
3. عامل الإنحدار الخطي.
4. تطبيق النموذج.
5. التحقق من صحة الأداء (الإنحدار الأداء).

- خطأ مطلق0.236 -/+ 0.346
- خطأ نسبي22.04% +/- 47.24%
- علاقة الإرتباط0.487

شجرة القرار

شكل رقم (4) أخذ عينات التدريب والاختبار



1. استرداد البيانات (بعد القاعدة).
2. تقسيم البيانات (0.5، 0.5).
3. عامل الإنحدار الخطي.
4. تطبيق نموذج.

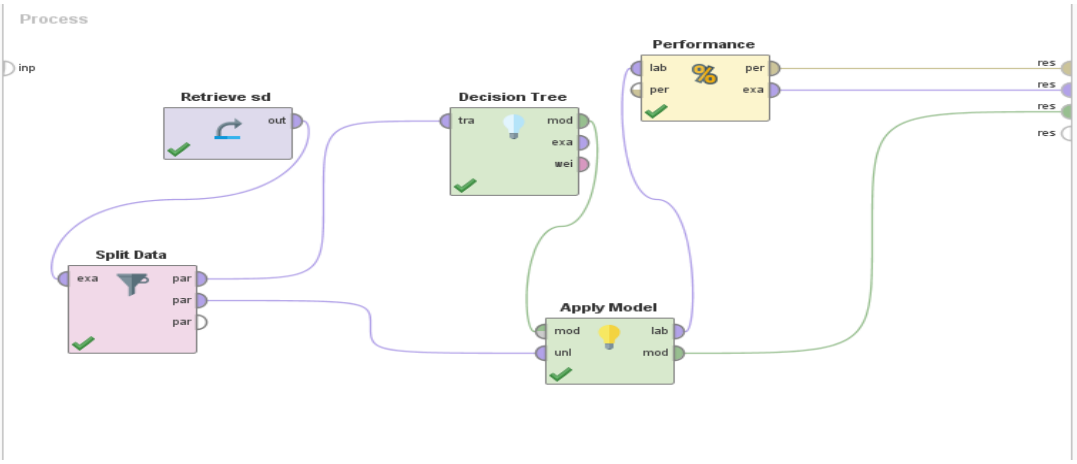
5. التحقق من صحة الأداء (الإنحدار الأداء).

الخطأ المطلق0.427-/+0.301

الخطأ النسبي.46.16% -/+ 50.50%

الإرتباط 0.340

شكل رقم (5) أخذ عينات التدريب والاختبار بعد الاختبار الأول



مقارنة أداء خوارزميات تنقيب البيانات المختلفة:

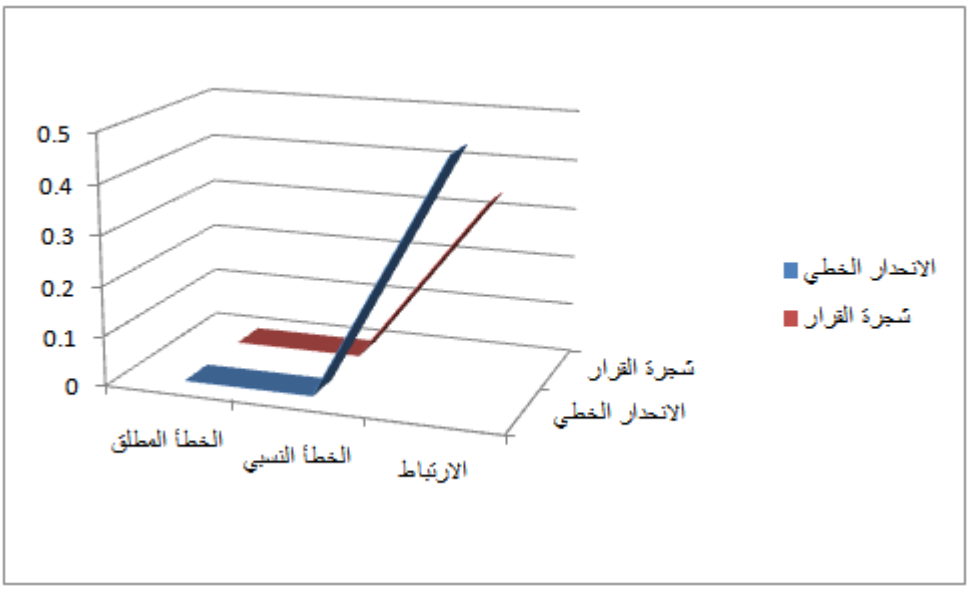
في هذا القسم سيعقد مقارنة أداء خوارزميات التنقيب عن البيانات التي يمكن أن تقوم بها التنبؤات، الإنحدار الخطي، الإنحدار الغير خطي، شجرة القرار ARAM أو ARIMA والشبكات العصبية في هذه الورقة سيتم مقارنة الإنحدار الخطي وشجرة القرار باستخدام Rapid Miner كما هو مذكور سابقا المعلمات التالية مفيدة للمقارنة:

1. الخطاء المطلق: أنه يمثل الإنحراف المطلق للتوقعات.
2. خطأ نسبي: يتم حساب الخطأ في القيمة المطلقة بين المتوقع للقيم والقيمة الحقيقية المعنية ويعبر عنه بالنسبة المئوية.
3. الإرتباط: أنه يوفر قيمة إرتباط مئوية بحد التنبؤ والقيم الفعلية في نطاق بحد 0 و100 حيث100 تمثل الكمال.

التنبؤ بالبيانات من خلال النموذج (يتم التعبير عنها بالنسبة المئوية)

نتائج العينة

مخطط رقم (2) نتائج العينة



التنبؤ الصحيح لشجرة القرار بال Rapid miner

جدول رقم (1) التنبؤ الصحيح لشجرة القرار بال Rapid miner

Open in Turbo Prep Auto Model Filter (241 / 384 examples): correct_predictions

Row No.	Type	predicti...	Pregna...	Glucose	BloodPr...	SkinThi...	Insulin	BMI	Diabete...	Age
1	0	0	1	89	66	23	94	28.100	0.167	21
2	0	0	5	116	74	0	0	25.600	0.201	30
3	0	0	10	139	80	0	0	27.100	1.441	57
4	1	1	1	189	60	23	846	30.100	0.398	59
5	0	0	1	103	30	38	83	43.300	0.183	33
6	0	0	8	99	84	0	0	35.400	0.388	50
7	1	1	11	143	94	33	146	36.600	0.254	51
8	0	0	5	109	75	26	0	36	0.546	60
9	0	0	3	88	58	11	54	24.800	0.267	22
10	1	1	9	102	76	37	0	32.900	0.665	46
11	0	0	1	146	56	0	0	29.700	0.564	29
12	0	0	7	105	0	0	0	0	0.305	24
13	0	0	1	101	50	15	36	24.200	0.526	26

التنبؤ الخاطى للإنحدار الخطى بال Rapid miner

جدول رقم (2) التنبؤ الخاطى للإنحدار الخطى بال Rapid miner

Open in Turbo Prep Auto Model Filter (384 / 384 examples): wrong_prediction

Row No.	Type	predicti...	Pregna...	Glucose	BloodPr...	SkinThi...	Insulin	BMI	Diabete...	Age
1	1	0.664	6	148	72	35	0	33.600	0.627	50
2	1	0.736	8	183	64	0	0	23.300	0.672	32
3	0	-0.066	1	89	66	23	94	28.100	0.167	21
4	0	0.168	5	116	74	0	0	25.600	0.201	30
5	0	0.626	10	115	0	0	0	35.300	0.134	29
6	1	0.707	2	197	70	45	543	30.500	0.158	53
7	0	0.860	10	139	80	0	0	27.100	1.441	57
8	1	0.790	1	189	60	23	846	30.100	0.398	59
9	1	0.650	5	166	72	19	175	25.800	0.587	51
10	1	0.382	0	118	84	47	230	45.800	0.551	31
11	0	0.397	1	103	30	38	83	43.300	0.183	33
12	1	0.299	1	115	70	30	96	34.600	0.529	32
13	0	0.375	8	99	84	0	0	35.400	0.388	50

اعتمادا على نتيجة الخطأ النسبي والإرتباط تبين أن شجرة القرار تتنبأ بنوع مرض السكرى أفضل من الإنحدار الخطي.

الخاتمة:

قامت هذه الورقة بدراسة إمكانية استخدام التنقيب عن البيانات للتنبؤ بمرض السكرى من النوع الأول والثانى. هنالك نماذج مستخدمة فى عملية التنبؤ بشكل عام وتم اختيار شجرة القرار والإنحدار الخطي وعمل مقارنة بينهم فى الخطأ المطلق والخطأ النسبي والإرتباط باستخدام Rapid Miner. من خلال الدراسة توصلنا إلى أن تطبيق التنقيب عن البيانات في التنبوء بمرض السكري يساعد ذلك في الإكتشاف المبكر مما يقلل من الآثار الجانبية وكذلك تقليل التكلفة الزمنية والمادية، بناءاً على هذه الدراسة يمكن استخدام طرق التنقيب عن البيانات لتحسين وضع التنبوء بمرض السكري في المؤسسات الصحية. كذلك توصلت الدراسة إلى نتيجة وهي أن نموذج شجرة القرار أفضل من نموذج الإنحدار الخطي.

النتائج:

1. تم تصميم إطا عمل للتنبؤ بمرض السكري وذلك باستخدام أدوات تنقيب البيانات.
2. الإكتشاف المبكر مما يقلل الأثار الجانبية.

3. قلة التكلفة الزمنية والمادية.
4. إتخاذ سبل الوقاية اللازمة للحد من إنتشار المرض.
5. تحسين وضع التنبؤ بمرض السكري في المؤسسات الصحية.
6. نموذج شجرة القرار أفضل من نموذج الإنحدار الخطي.

التوصيات:

1. جمع البيانات الحقيقية من المؤسسات الصحية بالسودان.
2. السجلات التي تم جمعها كانت 769 سجل، أوصت الباحثة بجمع عدد أكبر من السجلات للحصول على نتائج أفضل.
3. التوجه إلى أدوات تقنية ذكية تساعد في كشف وحلول علاج مرض السكري والوقاية منه.
4. التركيز على نماذج أخرى غير التي ذكرت في الدراسة.
5. التدريب والورش في مجال تنقيب البيانات خاصة مرضى السكري.
6. استخدام ال (deep learning) للتنبؤ بمرض السكري.

المصادر والمراجع

أولاً: المراجع العربية:

1. جلال إبراهيم العبد. استخدام الأساليب الكمية في إتخاذ القرارات الإدارية. دار الجامعة الجديدة.
2. حمدي طه. مقدمة في بحوث العمليات. دار المريخ.
3. نبيل محمد مرسي. الأساليب الكمية في الإدارة .

ثانياً: المراجع الأجنبية:

1. Cooper and Schindler (2003)Agrawal and Srikant ،1995 .
2. Cooper and Schindler ،2003.
3. Fu, Y., (1997), "Data Mining: Tasks, Techniques and Applications", Potentials, IEEE20–18 ,4 :16,.
4. Kantardzic, M., (2003), "*Data Mining: Concepts, Models, Methods, and Algorithms"* John Wiley and Sons, Inc., edited by ff.
5. SPSS White Paper, (1999)*, "Data mining: an introduction Clementine – Working with health care"*.

6. Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P., (1996a), "From data mining to knowledge discovery: an overview", *In* Fayyad, U., Shapiro, P. G., Smyth, P. and Uthurusamy, R. (eds.), *Advances in Knowledge Discovery and Data*, Palo Alto, Cambridge, Massachusetts: CA: AAAI/MIT Press, 1996, PP 1-34, also in *AI Magazine*, 17:3, 186-193.
7. SPSS White Paper, (1999)*, "Data mining: an introduction Clementine – Working with health care"*.
8. Read, B.J., (1999), "Data Mining and Science? Knowledge discovery in science as opposed to business", *12th ERCIM Workshop on Database Research.*
9. Ohsuga S., (2005), "Difference between Data Mining and Knowledge Discovery- A View to Discovery from Knowledge-Processing", *International Conference on Granular Computing*, Beijing, China, 25-27 July, *IEEE, 1*, 7-12.
10. Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P., (1996a), "From data mining to knowledge discovery: an overview", *In* Fayyad, U., Shapiro, P. G., Smyth, P. and Uthurusamy, R. (eds.), *Advances in Knowledge Discovery and Data*, Palo Alto, Cambridge, Massachusetts: CA: AAAI/MIT Press, 1996, PP 1-34, also in *AI Magazine*, 17:3, 186-193.
11. Read, B.J., (1999), "Data Mining and Science? Knowledge discovery in science as opposed to business", *12th ERCIM Workshop on Database Research.*
12. Fu, Y., (1997), "Data Mining: Tasks, Techniques and Applications", Potentials, IEEE.20–18 ,4 :16,
13. Mitra, S., and Acharya, T., (2003), Data Mining Multimedia, Soft Computing, and Bioinformatics, New Jersey, Hoboken, Published by John Wiley and Sons, Inc.
14. Han, J. and Kamber, M., (2006), Data Mining: Concepts and Techniques, San Francisco, U.S.A, Morgan Kaufman Publishers.
15. Han, J. and Kamber, M., (2006), Data Mining: Concepts and Techniques, San Francisco, U.S.A, Morgan Kaufman Publishers.
16. Langley, P., Simon, H., (1995), "Applications of Machine Learning and Rule Induction", Communications of the ACM", 38: 11, 54 – 64.
17. Hand, D. J., (1999), "Statistics and Data Mining: Intersecting Disciplines", ACM. SIGKDD Explorations Newsletter, 1: 1, 16 – 19.
18. Kantardzic, M., (2003), "*Data Mining: Concepts, Models, Methods, and Algorithms"* John Wiley and Sons, Inc., edited by ff.

ثالثاً: الشبكة الدولية للمعلومات

1. http://un.uobasrah.edu.iq/lectures/1988.pdf