بسم الله الرحمن الرحيم

جامعة إفريقيا العالمية

كلية الدراسات العليا

بحث لنيل درجة الماجستير في علوم الحاسوب

التنبؤ بمرض السكري وأنواعه باستخدام تنقيب البيانات

إعداد الطالبة:

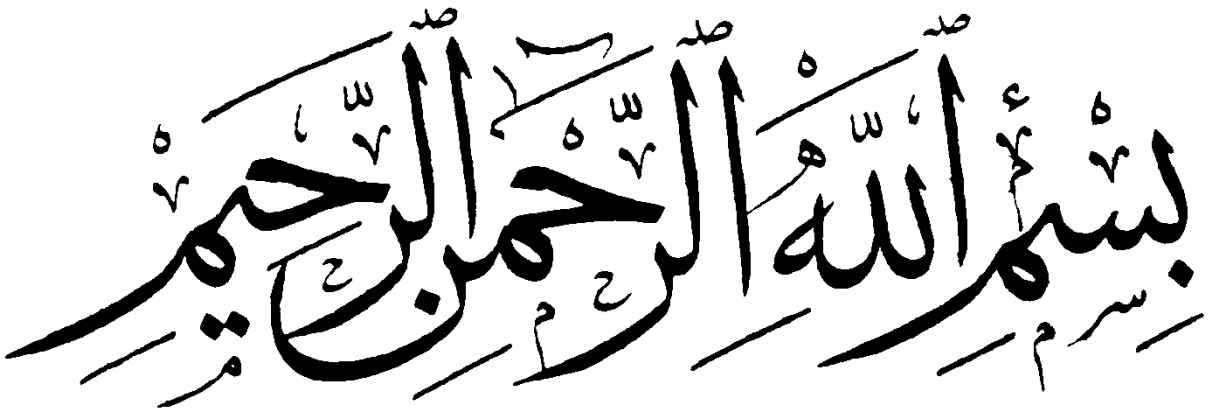
نسرين سامر عبد اللّه

إشراف :

أ. د. قسم السيد إبراهيم محمد

الخرطوم – السودان

(1442هـ - 2020م)


بسم الله الرحمن الرحيم

قَالَ تَعَالَىٰ:

﴿ قَالُوا۟ سُبْحَٰنَكَ لَا عِلْمَ لَنَآ إِلَّا مَا عَلَّمْتَنَآ ۖ إِنَّكَ أَنتَ ٱلْعَلِيمُ ٱلْحَكِيمُ ﴿٣٢﴾ ﴾

صدق اللّه العظيم

سورة البقرة: الآية 32

يا زهرة كلما تناثرت قطرات نداها شع بريقها في قلبي

يا زهرة أعطت للحياة معنى وللبسمة إشراق ما أنت عنكي قلت أو كتمت مع كل العبارات التي تلامس شفافيتك وصفاء روحك الطاهرة وقفت كلماتي عاجزة عن إعطائك حقك وليكن صمتي احتراماً ابلغ من كل العبارات.

أمي الحبيبة

جهراً أنادي باسمك المنسوج من برد التوهج والجمال حسبي لقائك في عيون الناس في بلدي جنوباً أو شمال سمر الملامح يشبهونك مشيةً أو قامةً لكنهم لا يشبهونك في الخصال

أبي الحبيب – رحمه الله

إلى من كانت ملاذي وملجئي

إلى من تذوقت معها أجمل اللحظات

إلى من سأفتقدها ..... وأتمنى أن تفتقدني

إلى من جعلها الله لي أخت في الله

صديقتي الوفية (رزان عبد الغفار)

وهم ينظرون بإشراق إلى الأيام القادمات فرحاً يقاوم أحزان الزمان وبشرى تحطم أسوار المستحيل، يظلون وقودي دائماً إلى تحقيق أي نجاح

أخواتي وإخواني

الآن تفتح الأشرعة وترتفع المرساة لتنطلق السفينة في عرض بحر واسع مظلم هو بحر الحياة وفي هذه الظلمة لا يضيء إلا قنديل الذكريات ذكريات الإخوة البعيدة

إلى الذين أحببتهم وأحبوني

زملائي وزميلاتي وكل من لهم الفضل في إكمال هذا البحث

الحمد لله رب العالمين والصلاة والسلام على أشرف الأنبياء والمرسلين سيدنا محمّد وعلى آله وصحبه ومن تبعهم بإحسان إلى يوم الدين، وبعد.

فإني أشكر الله تعالى على فضله حيث أتاح لي إنجاز هذا العمل بفضله، فله الحمد أولاً وآخراً.

ثم أشكر أولئك الأخيار الذين مدوا لي يدَ المساعدة، خلال هذه الفترة، وفي مقدمتهم أستاذي المشرف على الرسالة فضيلة الأستاذ الدكتور/ قسم السيد ابراهيم الذي لم يدَّخر جهداً في مساعدتي، كما هي عادته مع كل طلبة العلم، وكنت أجلس معه الساعات الطوال أقرأ عليه ولا يجد في ذلك حرجاً، وكان يحثّني على البحث، ويرغِّبني فيه، ويقوّي عزيمتي عليه فله من الله الأجر ومني كل تقدير حفظه الله ومتّعه بالصحة والعافية ونفع بعلومه.

المستخلص:

تكمن مشكلة البحث في التنبؤ بمرض السكري واستخدام التنقيب عن البيانات للتنبؤ بمرض السكرى من النوع الاول والثاني، أصبح التنقيب عن البيانات وتحليلها دراسة واسعة الانتشار في الآونة الأخيرة و يمكن تطبيقها على مجالات متنوعة حيث تستخرج هذه الطريقة عناصر بيانات غير محددة مسبقاً، في هذا البحث قام الباحث بدراسة إمكانية استخدام التنقيب عن البيانات للتنبؤ بمرض السكرى من النوع الاول والثاني، وتحديد الطريقة المناسبة للتنبؤ بمرض السكري باستخدام المنهج الوصفي التحليلي وذلك عن طريق التنقيب عن البيانات هناك نماذج مستخدمة في عملية التنبؤ بشكل عام سنختار منها شجرة القرار والانحدار الخطى و نقوم بعمل مقارنة بينهم في precision،accuracy ، Recall and F measure باستخدام Rapid Miner. استخدم الباحث بيانات (Pima Indians diabetics) التي تحتوي على 769 سجل و 9 خصائص .

عند تنفيذ خوارزمية الإنحدار الخطي داخل ال Rapidminer تحصلنا على (accuracy = 76.09 %) ، (precision = 79.14%) ، (Recall = 86.00%) و (F measure = 82.43%) و عند تنفيذ شجرة القرار تحصلنا على (accuracy = 70.87 %) ، (precision = 71.28%) ، (Recall = 92.67%) و (F measure = 80.58%) مع مقارنة النتائج التي تحصلنا عليها نجد أن الإنحدار الخطي أفضل من شجرة القرار في التنبوء بنوع مرض السكري

كلمات المفتاحية: التنقيب عن البيانات ، Rapidminer، شجرة القرار، الانحدار الخطي.

Abstract:

The research problem lies in predicting diabetes and using data mining to predict type 1 and type 2 diabetes. Data mining and analysis has become a widespread study in recent times and it can be applied to various fields where this method extracts unspecified data elements. The researcher is studying the possibility of using data mining to predict diabetes of the first and second types, and determining the appropriate method for predicting diabetes using the descriptive and analytical approach by mining the data. There are models used in the prediction process in general. We will choose from them the decision tree and the linear regression and make a comparison between them. In accuracy, precision, Recall and F measure using Rapid Miner. The researcher used the data (Pima Indians diabetics) that contain 769 records and 9 characteristics.

When executing the linear regression algorithm inside the Rapidminer، we get a

(accuracy = 76.09%)، (precision = 79.14%)، (Recall = 86.00%) and (F measure = 82.43%) and upon implementing the decision tree we got (accuracy = 70.87%)، (precision = 71.28%)، (Recall = 92.67%) and (F measure = 80.58%). By comparing the results we obtained، we find that linear regression is better than the decision tree in predicting the type of diabetes.

Keywords: data mining، rapidminer، decision tree، linear regression

فهرس الموضوعات

الفصل الأول

الإطار المنهجي

الفصل الأول

الاطار المنهجي

1.1 المقدمة

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 استعمالاتها أوفي طرق اتخاذ القرار ومن هذه الأدوات التنقيب عن البيانات.

التنقيب هو عمليه بحث محوسب ويدور عن معرفة من البيانات دون فرضيات مسبقه عما يمكن ان تكون هذه المعرفة. ويعرف التنقيب في البيانات على انه عملية تحليل كمية بيانات، لإيجاد علاقه منطقيه تلخص البيانات بطريقه جديده تكون مفهومه ومفيدة.

ظهر التنقيب في البيانات في اواخر الثمانيات واثبت وجوده كأحد الحلول لتحليل كميات ضخمه من البيانات وذلك بتحويلها من مجرد معلومات متراكمه وغير مفهومه الى معلومات قيمه يمكن استغلالها والاستفادة منها، كما يعرف انه تحليل كميه من البيانات عاده ما تكون كبيره لإيجاد علاقه منطقيه تلخص البيانات بطريقة جديدة تكون مفهومه ومفيدة.

صاحبت الطفرة التي حدثت في تطوير أجهزة وأدوات وأساليب التشخيص، ثورة أخرى في تقدم كثير من برامج الحاسب الآلي، من أجل توفير سرعة رصد وتحليل النتائج، وكان لابد من هذا الجانب لتلبية متطلبات هذه التقنيات الجديدة.

واستطاع الباحثون والعلماء، استخدام الكمبيوتر في تطوير أجهزة التشخيص والحصول على نتائج فورية دقيقة، إضافة إلى ابتكار بعض التقنيات والاختبارات الجيدة التي تفيد في التوصل إلى تشخيص بعض الأمراض بطرق أكثر دقة وسرعة.

توصل العلماء من خلال هذه التطورات إلى طريقة تحليل الشعر، التي تعطي نتائج أكثر دقة عن وجود المعادن والأملاح بالجسم، وبذلك تخلص الأطباء من أخطاء وخلل الطرق التقليدية لمعرفة نسب المعادن في الجسم، والتي غالباً ما تكون غير دقيقة في مجمل أنواعها.

كما نجح الباحثون في تطوير تقنية للاختبارات الجينية التي ستصبح أكثر حساسية، وتتجاوز مشكلة التغطية على الجين المعطوب بواسطة الجين السليم، وستحدث ثورة للكشف عن الأمراض الوراثية.

وتم التوصل إلى اختبارات جديدة تفيد في كشف مرض الزهايمر مبكراً، ومن ثم اتخاذ التدابير اللازمة لإيقاف عملية تدمير الدماغ. وفي هذا الموضوع سنعرض التطورات والتقنيات الجديدة والحديثة في معرفة بعض البيانات، وتطوير طرق الفحص ومعرفة عدد من الأمراض والتخلص من مشاكل الطرق الحالية والتقليدية في الاختبارات والتحليلات .

في هذا البحث سوف تكون دراسة الحالة مرض السكري وسوف نقوم بالتنقيب عن البيانات للتنبؤ بمرض السكري من النوع الأول و النوع الثاني

2.1 مشكلة البحث:

تكمن مشكلة البحث في التنبؤ بمرض السكري واستخدام التنقيب عن البيانات للتنبؤ بمرض السكرى من النوع الاول والثاني، كذلك مساعدة الباحثين في اختيار التقنية المناسبة للتنبؤ بالمرض

3.1 أهداف البحث:

يتمثل الهدف الرئيسي للبحث في إيجاد طريقة معينة للتنبؤ بمرض السكري باستخدام تقنيات التنقيب عن البيانات مما ينتج أهداف فرعية تتمثل في دراسة مشكلة التنبؤ بمرض السكري، وللتحقق من امكانية استخدام أدوات التنقيب عن البيانات للتنبؤ بمرض السكري، كذلك تحديد طريقة مناسبة للتنبؤ بمرض السكرى.

4.1 أسئلة البحث:

1. ما هي الأسباب التي تساعد في التنبؤ بمرض السكري؟.
2. كيف يتم استخدام تنقيب البيانات للتنبؤ بمرض السكري؟
3. هل يساعد تنقيب البيانات في التقليل من الوقت المستهلك في التنبؤ بالمرض؟

5.1 أسباب اختيار الموضوع:

اختارت الباحثة موضوع البحث وذلك لأهميته وانتشار مرض السكري بصورة كبيرة في الآونة الأخيرة.

يساعد التنقيب في التنبؤ بالمرض في مناطق محدده او ازمنه معينه او ظروف واحوال بعينها، بهدف وضع الحلول المناسبة واتخاذ سبل الوقاية اللازمة للحد من انتشار المرض.

6.1 أهمية البحث:

تكمن أهمية البحث في أهمية موضوع مرض السكري وذلك بـــ:

1. مرض السكري من أكثر الأمراض انتشاراً وقد تكون أمراض وراثية.
2. مرض السكري من الأمراض المزمنة وينتج عنه بعض التأثيرات الجانبية (النظر، الأسنان ، الكلى و بتر الأطراف)
3. الاستفادة من البيانات الموجودة لدى المؤسسات الصحية في التنبؤ بنوع مرض السكري النوع (الاول ،الثاني)
4. تقليل التكلفة المادية والزمنية في التنبؤ بنوع مرض السكري.

7.1 فروض البحث:

1. تنقيب البيانات يساعد علي اكتشاف علاقات جديده من واقع البيانات المخزنة.
2. تنقيب البيانات يساعد على استغلال البيانات بصوره جيده .

8.1 منهجية البحث:

هذا البحث يتبع منهج وصفى تحليلي، تم تنفيذ هذا المنهج في data set وتم جمعها عن طريق الانترنت تحتوي على تسعة خصائص وهي: Pregnancies، Glucose، blood pressure، skin thickness، Insulin ، BMI ، Function Pedigree Diabetes ، Age، Outcome وتنفيذ نموذج شجرة القرار ونموذج الانحدار الخطي.

9.1 وسائل وأدوات البحث:

ادوات التنقيب عن البيانات مثل weka ، rapidminer واستخدام النماذج مثل (شجرة القرار ، الانحدار الخطي ، الشبكات العصبية ...إلخ)

10.1 حدود البحث:

- الحدود المكانية :السودان – الخرطوم.
- الحدود الزمانية 2018م– 2020م

11.1 هيكل البحث: يحتوي البحث على خمسة فصول حيث يتناول الفصل الأول، الإطار المنهجي والدراسات السابقة وذلك من خلال مبحثين، المبحث الأول: الإطار المنهجي، والمبحث الثاني: الدراسات السابقة.

الفصل الثاني: الإطار النظري والذي يحتوي على المصطلحات العلمية واختيار نموذج التنقيب عن البيانات وينقسم إلى مبحثين، المبحث الأول: مرض السكري، المبحث الثاني: التنقيب عن البيانات

الفصل الثالث: المنهجية والطرق المستخدمة في هذا البحث

الفصل الرابع: يحتوي على العمليات المسبقة للData Mining في طريقة التنفيذ

الفصل الخامس: يحتوي على نتائج التحليل والخاتمة والتوصيات.

المبحث الثاني الدراسات السابقة

اطلع الباحث على العديد من الدراسات السابقة المعنية بالمجال والتي تناولت ذات الموضوع بحثا، وتيسر للباحث ذلك من خلال اطلاعه في بعض المكتبات فيها بالضرورة مكتبة جامعه افريقيا العالمية ، وقد تم ايجاد بعضا من البحوث التي تطرقت لجزئيات تشير عن ذات الموضوع وهي كما يلي :

- الدراسة الأولى: Badge،R ، and p. Patil ، Diagnosis of Diabetes using OLAP and data mining integration . International journal of computer science &communication networks،[1].322-314.p:(3)2012.2 ، قدما نظام دعم القرار والذي يجمع بين نقاط القوة في كلا OLAP واستخرج البيانات بحيث يمكن الأطباء من التنبؤ بالمرضى الذين يتم تشخيصهم بالإصابة بداء السكري. لخصت الدراسة أيضا إلى أن العلاج عن طريق العقاقير للمرضى في الفئة العمرية المبكرة من الشباب يمكن أن تتأخر لتجنب الأثار الجانبية وإفساح المجال لأساليب العلاج الأخرى مثل الرياضة . في المقابل ، فإن المرضى في الفئة العمرية المتقدمة من العمر يجب أن يوصف العلاج لهم عن طريق العقاقير.

- الدراسة الثانية: Concaro،s. ، l. sacchi ، and R. Bellazzi. Temporal data mining for the assessment of the costs related to diabetes mellitus pharmacological treatment. In AMIA 2009 Symposium proceedings.200912] قدم تطبيق تقنية التنقيب عن البيانات الزمنية لاستخراج القواعد النقابية الزمنية على مستودع متكامل على حد السواء بما في ذلك البيانات الإدارية والسريرية المتعلقة من عينة مرضى السكري.

- الدراسة الثالثة: Mukwevho، p. ، evaluation of health and diabetes knowledge of cu 4 health participants by food frequency questionnaire and Michigan diabetes knowledge test، in food ، nutrition and culinary sciences. 2010، the graduate school of Clemson university ، في دراسه استنتج

أن التدخل في تعديل نمط الحياة قد اثبت أن يكون وسيلة فعالة لعلاج وتأخير مرض السكري من النوع التانى الذي ينشأ من زيادة الوزن. وعدم ممارسة الرياضة.

- الدراسة الرابعة: DATA MINING Technique in diabetes، Journal of ، Anita Shaikh and Sohail Abul Sattar،Saman Hina' Department of ، 466-471، 13، 2017،Basic & Applied Sciences NED University of ،Computer Science & Software Engineering Karachi Pakistan،Engineering & Technology

استخدام تقنية تنقيب البيانات في مرض السكري للاستفادة من البيانات الكبيرة المخزنة دون الاستفادة منها.

واستنتجت الدراسة تحسين التشخيص الطبي لمرض السكري والحد من الوقت.

- الدراسة الخامسة: Analysis Diabetes using data mining

في دراسة تحليل مرضى السكري من النوع الثاني باستخدام تنقيب البيانات، مركز الخرطوم للعلاج بالأشعة والطب النووي.

إعداد الطالبة: أم كلثوم صباحي محمد حمدون، إشراف الأستاذ: وليد احمد خلف الله، سبتمبر 2011م.

استنتجت الدراسة:

- أنه الأكثر شيوعاً في جميع أنحاء العالم وتنبأت الدراسة بأنه سوف تصل إلى 12،000 حالة بحلول العام 2035م
- استخدمت أداة Wekaفي استخراج البيانات
- الدراسة السادسة: R. Sivanesan K. Devika Rani Dhivya، *A Review on Diabetes Mellitus diagnoses using classification on Pima Indian Diabetes Data Set*، International Journal of Advance

- تتضمن هذه الورقة تشخيص مرض السكري باستخدام تقنية التصنيف ، وهي تقنية تستخدم بشكل عام في استخراج البيانات الطبية. وتعتبر عملية إيجاد دالة أو نموذج لتقسيم البيانات إلى فئات مختلفة من كائن بناءً على خصائصه. على عكس نماذج ونهج التصنيف المستخدمة لتشخيص وعلاج مرض السكري. تركز هذه الورقة على أداء(شجرة القرار) J48 Decision Tree ، وهي نموذج تنبؤي لتعلم الآلة يتبنى القيمة المستهدفة (المتغير التابع) لعينة جديدة بناءً على قيم السمات المتنوعة للبيانات التي يمكن الوصول إليها. يتبع مصنف شجرة القرار J48 الخوارزمية البسيطة. من أجل تصنيف عنصر جديد ، يجب أولاً إنشاء شجرة قرارات بناءً على قيم السمات لبيانات التدريب المتاحة. تتكون مجموعة البيانات الخاصة بمرضى السكري المأخوذة من مستودع التعلم الآلي من 768 حالة مع 9 سمات. يتم تقييم مجموعة البيانات باستخدام مجموعة التدريب ، والتحقق المتقاطع من 10 أضعاف وطريقة تقسيم النسبة المئوية والنتائج التي تبني عليها الخوارزمية أفضل النماذج بطريقة فعالة.

- الدراسة السابعة: Nandana Pathirage and ،L.H.S De Silva# ،T.M.K.K Jinasena ‘Diabetic Prediction System Using Data Mining، *Department of Computer Science*، *Faculty of Applied Sciences*، *University of Sri Jayewardenepura*، *2016*

- تستخدم هذه الورقة العلمية استخراج البيانات حيث يعد استخراج البيانات أحد المجالات الرئيسية للتعلم الآلي. يلعب دوراً مهماً في أبحاث مرض السكري لأنه يمتلك القدرة على استخراج المعرفة الخفية من كمية كبيرة من البيانات المتعلقة بمرض السكري. الهدف من هذا البحث هو تطوير نظام يمكنه التنبؤ بما إذا كان

المريض يعاني من مرض السكري أم لا. علاوة على ذلك ، فإن التنبؤ بالمرض في وقت مبكر يؤدي إلى علاج المرضى قبل أن يصبح حرجًا. ركز هذا البحث على تطوير نظام قائم على ثلاث طرق تصنيف وهي شجرة القرار ، Naïve Bayes وخوارزميات آلة ناقلات الدعم. حاليًا ، تعطي النماذج دقة 84.6667٪ و 76.6667٪ و 77.3333٪ لشجرة القرار و Naïve Bayes و SMO Support Vector Machine على التوالي. تم التحقق من هذه النتائج باستخدام منحنيات خصائص تشغيل جهاز الاستقبال بطريقة حساسة للتكلفة. تستخدم طريقة المجموعة المتقدمة الأصوات التي أعطتها الخوارزميات الأخرى للحصول على النتيجة النهائية. تلغي آلية التصويت هذه التصنيفات الخاطئة المعتمدة على الخوارزمية. تظهر النتائج تحسناً كبيراً في دقة طريقة المجموعة مقارنة بالطرق الأخرى

- الدراسة الثامنة: N. Snehaand Tarun Gangil، Analysis of diabetes mellitus for early prediction using optimal features selection، Sneha and Gangil، journal of Big Data، 2019.

- تهدف هذه الدراسة من الاستفادة من الميزات المهمة ، وتصميم خوارزمية التنبؤ باستخدام التعلم الآلي والعثور على المصنف الأمثل لإعطاء أقرب نتيجة مقارنة بالنتائج السريرية. تهدف كذلك إلى التركيز على اختيار السمات التي تتسبب في الكشف المبكر عن مرض السكري Miletus باستخدام التحليل التنبئي. تظهر النتيجة خوارزمية شجرة القرار والغابة العشوائية لها أعلى خصوصية تبلغ 98.20٪ و 98.00٪ ، على التوالي هي الأفضل لتحليل بيانات مرضى السكري. نتائج Naïve Bayesian تشير إلى أفضل دقة 82.30٪. يعمم البحث أيضًا اختيار الميزات المثلى من مجموعة البيانات لتحسين دقة التصنيف.

الرقم	اسم الباحث	عنوان البحث	النتائج	–
1	R،Badge ، and p. patil	Diagnosis of Diabetes sing OLAP and data miningintegration	نظم دعم القرار والذى يجمع نقاط القوة فى كلا olap واستخرج الببيانات بحيث يمكن الاطباء من التنبؤ بالمرض	من توصيات هذا البحث المرضى فى الفئة العمرية والمتقدمة من العمر يجب ان يوصف العلاج لهم عن طريق العقاقير واستخدام الاطباء التقنيات الذكية والتعلم والمهارات عليها.
2	Concaro، s.، l. sacchi ، and R. Bellazzi	Temporal data mining for the assessment of the costs related to diabetes mellitus pharmacological treatment	قدم تطبيق تقنية التنقيب عن البيانات الزمنية لاستخراج القواعد النقابيه الزمنية على مستودع متكامل	العمل على تدريب الاطباء استخدام هذه التقنيات.
3	Mukwevho ،p	evaluation of health and diabetes knowledge of cu 4 health participants by food frequency questionnaire and Michigan diabetes knowledge test ، in	التدخل فى تعديل نمط الحياة قد اثبتت ان يكون وسيلة فعالة لعلاج وتاخير مرض السكرى من النوع التانى الذى ينشأ من زيادة الوزن. وعدم ممارسة الرياضة.	ايجاد طرق للكشف المبكر لهذا المرض خصوصا وانه وراثى اى يمكن للشخص معرفة ان يكون متوقع حدوث المرض له منذ الطفولة عن طريق الكشف.

		food ، nutrition and culinary sciences		
4	Saman Hina’ ، Anita Shaikh and Sohail Abul Satta	DATA MINING Technique in diabetes	تحسين التشخيص الطبى لمرض السكرى والحد من الوقت.	عمل ورش ودورات للاطباء فى استخدام ادوات تقنيات تنقيب البيانات.
5	أم كلثوم صباحي محمد حمدون	Analysis Diabetes using data mining	وتنبأت الدراسة بانه سوف تصل الى 12.000 حالة بحلول العام2035م	ايجاد اجهزة حديثة ومهارات للكشف المبكر لهذا المرض.
6	R. Sivanesan K. Devika Rani Dhivya	A Review on Diabetes Mellitus diagnoses using classification on Pima Indian Diabetes Data Set	من أجل تصنيف عنصر جديد ، يجب أولاً إنشاء شجرة قرارات بناءً على قيم السمات لبيانات التدريب المتاحة	التحقق المتقاطع من طريقة تقسيم النسبة المئوية والنتائج التي تبني عليها الخوارزمية
7	L.H.S De Silva#، Nandana Pathirage and T.M.K.K Jinasena	Diabetic Prediction System Using Data Mining	تلغي آلية التصويت هذه التصنيفات الخاطئة المعتمدة على الخوارزمية. تظهر النتائج تحسنًا كبيرًا في دقة طريقة المجموعة مقارنة بالطرق الأخرى	إستخدام طريقة المجموعة في بحوث التنبوْ يإستخدام التنقيب عن البيانت و إجراء المزيد من المقارنات

الفصل الثاني

الإطار النظري

الفصل الثاني

المبحث الأول

مرض السكري

1.2 تعريف مرض السكري وأنواعه:

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

وهرمون الأنسولين له فاعلية أساسية في عمليات الاستقلاب والتعامل مع الغذاء بشكل عام و مع 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 يتقلب في إناث اختلالات عديدة قد تظهر على المدى القريب أو البعيد.

ويندرج تحت ما يسمي 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.

1- السكري من النوع الأول (Diabetes Mellitus Type 1)

هو حالة مزمنة ينتج فيها البكرياس كمية صغير من الانسولين

2- السكري من النوع الثاني (Diabetes Mellitus Type 2)

اضطراب استقلابي يتميز بارتفاع معدل السكر في الدم في سياق مقاومة الانسولين ونقص الانسولين النسبي. يشكل السكرى من النوع الثاني 90% من حالة مرض

السكري ويعتقد ان السمنة هي السبب الرئيسي لسكري النمط الثاني واعراضه الكلاسيكية هي العطش الزائد وكثرة التبول و شعور ومتواصل بالجوع.

الشعار العالمي لمرض السكري

أنواع أخرى:

أ) السكري الناتج عن بعض أمراض البنكرياس.

ب) السكري الناتج عن اختلالات هرمونية وخصوصا في الغدد النخامية والكظرية وخلايا (1) في البنكرياس.

ج) السكري الناتج عن بعض الأدوية.

د) أنواع أخرى نادرة.

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

(1) المرضى ذوو الاحتمالات الكبيرة جداً للمضاعفات الخطيرة بصورة مؤكدة طبيا وتتميز أوضاعهم المرضية بحالة أو أكثر مما يأتي:

- حدوث هبوط السكر الشديد خلال الأشهر الثلاثة التي سبقت شهر رمضان.
- المرضى الذين يتكرر لديهم هبوط وارتفاع السكر بالدم.
- المرضى المصابون بحالة (فقدان الإحساس بهبوط السكر)، وهي حالة تصيب بعض مرضى السكري، وخصوصا من النوع الأول، الذين تتكرر لديهم حالات هبوط السكر الشديد ولفترات طويلة.

- المرضى المعروفون بصعوبة السيطرة على السكري لفترات طويلة.
- حدوث مضاعفة (الحماض السكري الكيتوني) أو مضاعفة (الغيبوبة السكرية الأسمولية) خلال الشهور الثلاثة التي سبقت شهر رمضان.
- السكري من النوع الأول.
- الأمراض الحادة الأخرى المرافقة للسكري.
- مرضى السكري الذين يمارسون مضطرين لأعمال بدنية عنيفة.
- مرضى السكري الذين يجري لهم غسيل كلى.
- المرأة المصابة بالسكري أثناء الحمل.

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

- الذين يعانون من ارتفاع السكر في الدم كأن يكون المعدل 180 – 300مغم/ دسل، (10ملم –5. 16 ملم) ونسبة الهيموغلوبين المتراكم (المتسكر) التي تتجاوز 10% .
- المصابون بقصور كلوي.
- المصابون باعتلال الشرايين الكبير (كأمراض القلب والشرايين).
- الذين يسكنون بمفردهم والذين يعالجون بواسطة حقن الأنسولين أو العقارات الخافضة.
- الذين يعانون من أمراض أخرى تضيف أخطارا إضافية عليهم.
- كبار السن المصابون بأمراض أخرى مثل السرطان.
- 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.

و بعد دراسة هذه الحالات من المرض أو الأقسام الأربعة من المرضى و مناقشتها بين الفقهاء والأطباء و ملاحظة الفروق الدقيقة بينها خلال الندوة الخاصة التي عقدتها المنظمة الإسلامية للعلوم الطبية بدولة الكويت في عام 2007 عن هذا الموضوع توصل المشاركون إلى قرار عن الحكم الشرعي لقيام المصاب بمرض السكري بالصيام أو تركه و تأجيله لأيام أخر و هذا نص القرار الذي اتخذته الندوة.

المبحث الثاني

التنقيب عن البيانات

2.2 نظرة عامة على تنقيب البيانات

"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 اقتراح تقنية متعددة التخصصات تسمى تنقيب البيانات.

في الواقع ، "الحاجة إلى فهم مجموعات البيانات الكبيرة والمعقدة الغنية بالمعلومات أمر شائع تقريبًا في جميع مجالات الأعمال والعلوم والهندسة. في عالم الأعمال ، أصبحت بيانات الشركات والعملاء معروفة كأصل استراتيجي الإجراء الكامل لاستخدام أسلوب قائم على الكمبيوتر ، إلى جانب تقنيات جديدة ، لاكتشاف المعرفة والمعلومات من البيانات يسمى بـــ "تنقيب البيانات".

يعد التنقيب عن البيانات مفيداً للغاية في سيناريو التحليل الاستكشافي الذي لا توجد فيه مفاهيم محددة مسبقًا حول ما الذي سيشكل نتيجة" مثيرة للاهتمام في سياق الأعمال، يعد تنقيب البيانات عملية تحديد الميزات أو العلاقات أو الأنماط أو النماذج المثيرة للاهتمام من قواعد البيانات الكبيرة من أجل إدارة أعمالك بشكل أفضل.

يعد التنقيب عن البيانات هو الجزء الأساسي والجزء الرئيسي من اكتشاف المعرفة في قاعدة البيانات. يحتوي الإجراء KDD عادة على الخطوات التالية: جمع البيانات ، تصحيح البيانات ، تحويل البيانات ، استكشاف الأنماط (تنقيب البيانات) ، تفسير النتائج ، التقييم واستخدام المعرفة المكتشفة.

"العديد من البائعين والمستشارين والمحللين يجعل تنقيب البيانات يبدو معقداً وصعباً مكلفًا. قد يكون الأمر في بعض الأحيان معقداً (يتضمن العديد من الأجزاء) ، لكن لا يلزم أن يكون غامضاً أو صعباً تنقيب البيانات يعني ببساطة: العثور على أنماط في البيانات الخاصة بك والتي يمكنك استخدامها لإجراء أعمالك بشكل أفضل

"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 من الإجابات. يمكن التحقق من صحة المجاملات التي تم تحقيقها من خلال تنقيب البيانات من خلال التحليل التقليدي.

3.2 الفرق بين اكتشاف المعرفة وتنقيب البيانات

كل من التنقيب عن البيانات واكتشاف المعرفة هي تقنيات للحصول على بعض سمات الكيان من مجموعة من البيانات على مكون فردي إلى كل من الخصائص / الميزات التي تم الحصول عليها من قبل المراقبة.

يشير مصطلح "اكتشاف المعرفة في قواعد البيانات" إلى العملية الكاملة لاكتشاف المعرفة القيمة من البيانات. إن اكتشاف المعرفة في قواعد البيانات هو عملية التعرف على الأنماط / النماذج المناسبة والرواية والتي يمكن أن تكون عملية، وفي النهاية مفهومة في البيانات.

يعد (تنقيب) البيانات العنصر الأساسي في الإجراء الأكثر شيوعاً لاكتشاف المعرفة في قواعد البيانات.

4.2 أوصاف اكتشاف المعرفة وتنقيب البيانات من الدراسات

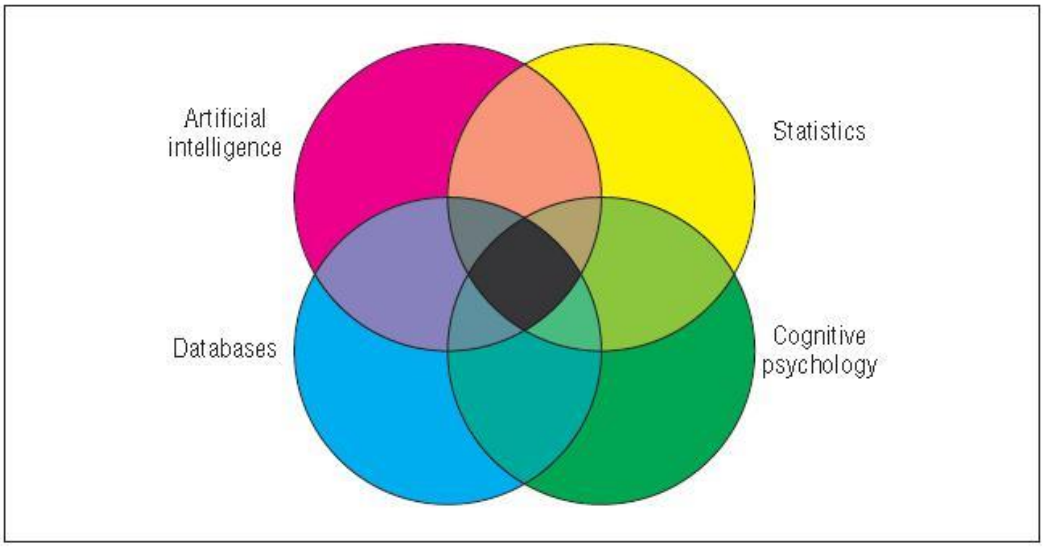
"التنقيب عن البيانات هو العملية غير التوفيقية لتحديد أنماط البيانات الصحيحة والجديدة والتي يمكن أن تكون مفيدة ويمكن فهمها في النهاية"

"يعد تنقيب البيانات مجالًا متعدد التخصصات يجمع بين تقنيات التعلم الآلي والتعرف على الأنماط والإحصائيات وقواعد البيانات والتصور لمعالجة مسألة استخراج المعلومات من قواعد البيانات الكبيرة.

"يتيح تنقيب البيانات ، والمعروف أيضاً باكتشاف المعرفة في قواعد البيانات ، للمؤسسات الأدوات اللازمة لتفحص مخازن البيانات الضخمة هذه للعثور على

الاتجاهات والأنماط والعلاقات التي يمكن أن توجه عملية اتخاذ القرارات الاستراتيجية (Ganti et al. 1999). "،

KDD"هي الخطوة العملية الأولى نحو تحقيق المعلومات كعامل إنتاج؛ استشهد به. ذكرت Pazzani في عام 2000 أن "حقل اكتشاف المعرفة وتنقيب البيانات (KDD) يعتمد على النتائج المستخلصة من الإحصاءات وقواعد البيانات والذكاء الاصطناعي لبناء الأدوات التي تتيح للمستخدمين اكتساب نظرة ثاقبة من مجموعات البيانات الضخمة" (الشكل 6).



الشكل 2: الجمع بين علم النفس المعرفي والذكاء الاصطناعي وقواعد البيانات والإحصاءات، المصدر: **Pazzani)،(2000**

"تنقيب البيانات هو استكشاف وتحليل كميات كبيرة من البيانات من أجل اكتشاف أنماط وقواعد ذات معنى.

"مجال التنقيب عن البيانات مكرس لتحليل البيانات للعثور على الاتصالات الأساسية واكتشاف أنماط جديدة

"تنقيب البيانات هو عملية اكتشاف معرفة مثيرة للاهتمام من كميات كبيرة من البيانات المخزنة إما في قواعد البيانات أو مستودعات البيانات أو مستودعات المعلومات الأخرى.

يشير تنقيب البيانات ببساطة إلى استخراج المعرفة أو "استخراجها" من كميات كبيرة من البيانات.

5.2 عملية تنقيب البيانات واكتشاف المعرفة

في الواقع (KDD) هو إجراء تكراري وتفاعلي يتضمن مختلف المراحل (فياض وآخرون، 1996 ب) والتي تم عرضها من قبل في الشكل 5 والموضحة بالتفصيل أدناه:

1.5.2 تحديد المشكلة وفهم مجال التطبيق

والخطوة الأساسية هي لفهم مجال التطبيق. من الواضح أن هذه الخطوة هي شرط مسبق لاستخراج المعرفة القيمة واختيار تقنيات تنقيب البيانات المناسبة في الخطوة الثالثة وفقًا لهدف التطبيق وطبيعة البيانات.

2.5.2 بيانات الاختيار والمعالجة المسبقة

"الخطوة الثانية هي جمع ومعالجة البيانات مسبقًا، يلزم إجراء معالجة مسبقة للبيانات لتحسين جودة البيانات الحقيقية للتنقيب، البيانات التي نريد تحليلها بواسطة طرق تنقيب البيانات غير كاملة وصاخبة وغير متناسقة ؛ وبالتالي ، هناك حاجة إلى معالجة البيانات المسبقة الخطوات الأساسية لإعداد البيانات موصوفة أدناه

تصحيح البيانات: يتكون من بعض العمليات الأساسية ، مثل تحديد القيم المتطرفة وتصحيح التناقضات في البيانات وتطبيعها وحذف الضوضاء ومعالجة البيانات المفقودة وتقليل التكرار وما إلى ذلك بيانات العالم الحقيقي غير صحيحة بشكل منتظم وغير كاملة وغير متناسقة ، ربما بسبب النظام عيوب الأداء ، خطأ بشري أو تشغيلي. تتطلب البيانات منخفضة الجودة وتصحيحها قبل تنقيب البيانات تكامل

البيانات وتحويلها: تحتوي هذه العملية على دمج العديد من قواعد البيانات غير المتجانسة المنتجة من مصادر متنوعة علاوة على ذلك ، قد تحتاج البيانات إلى تحويلها إلى أشكال مناسبة للتنقيب.

تقليل البيانات: يمكن تطبيق التقنيات للحصول على تمثيل مخفض لحجم مجموعة البيانات ؛ ومع ذلك ، يحتفظ تكامل البيانات الأصلية. وهذا يعني أن التنقيب على مجموعة البيانات المنخفضة يجب أن يكون أكثر كفاءة ، لكنه يخلق النتائج التحليلية نفسها أو تقريباً.

تقديرية البيانات: يمكن تطبيق التقنيات لتقليل عدد القيم لميزة مستمرة معينة عن طريق فصل سلسلة الخاصية إلى فواصل زمنية. بعد ذلك ، يمكن تطبيق علامات الفاصل الزمني لاستبدال قيم البيانات الحقيقية.

1. تنقيب البيانات:

والخطوة الثالثة هي تنقيب البيانات التي تستخرج الأنماط الخفية والمعرفة المفيدة. يعد هذا إجراءً حاسماً حيث تستخدم التقنيات الذكية لاستخراج المعرفة والأنماط من البيانات

2. نتيجة التفسير والتقييم:

تتضمن هذه الخطوة تفسير المعرفة التي تم تحقيقها ، خاصةً التفسير من حيث الوصف أو التنبؤ ، وهما هدفان رئيسيان لتنفيذ أنظمة الاكتشاف

3. تطبيق المعرفة المكتشفة:

إن وضع النتائج في الاستخدام العملي هو بالتأكيد الهدف النهائي لاكتشاف المعرفة يمكن تطبيق المعلومات أو المعرفة التي تم الحصول عليها بواسطة تقنيات تنقيب البيانات لاحقًا لتنوير الاتجاهات أو الحقائق الموجودة أو التاريخية ، والتنبؤ بالمستقبل، ومساعدة صناع القرار على وضع استراتيجية من الحقائق والمعلومات المستخرجة.

2.1.5 بيانات آلة التعلم والإحصاء

التنقيب عن البيانات لا يحل محل الأساليب الإحصائية التقليدية. إنها امتداد للتقنيات الإحصائية التي تعد جزئيًا نتيجة لتغيير رئيسي في مجال الإحصاء كان نمو معظم الطرق التقليدية وفقًا للنظرية الذكية والتقنيات التحليلية التي عملت جيدًا نسبيًا على الكميات المختلفة من البيانات التي يتم فحصها.

وفقًا لقدرة الحاسوب المحسّنة على الكم الهائل من البيانات التي يمكن الوصول إليها ، يمكن للطرق الجديدة تقدير تقريباً أي مخطط أو تفاعل وظيفي غير مصحوبين تعتمد الأساليب الإحصائية التقليدية على المصمم لتحديد الشكل الوظيفي والتفاعلات النقطة الأساسية هي أن التنقيب عن البيانات هو وظيفة الذكاء الاصطناعي والأساليب الإحصائية لقضايا العمل العامة بحيث تجعل هذه الأساليب في متناول فاحص المعرفة ذي الخبرة بالإضافة إلى خبير الإحصاء الماهر.

6.2 تنقيب البيانات والتعلم الآلي

يتم تطبيق التعلم الآلي على التقنيات الحسابية التي تهدف إلى زيادة الأداء من خلال أتمة اكتساب المعرفة والمعلومات من التجربة

.يهدف التعليم الآلي إلى تقديم مستويات متزايدة من الميكنة في إجراء هندسة المعرفة، والاستعاضة عن النشاط البشري الذي يستغرق وقتًا طويلًا بطرق تلقائية تعمل على تحسين الدقة أو الكفاءة من خلال اكتشاف وتطبيق عناصر انتظامية في بيانات التدريب.

7.2 تنقيب البيانات والإحصاء

تنطوي الإحصائيات وتنقيب البيانات كلاهما اكتشاف بنية البيانات نظراً لتداخل كميات كبيرة من خططهم ، يعتبر بعض الأشخاص تنقيب البيانات تقسيماً للإحصاءات ومع ذلك ، فإن هذا ليس تقييماً عقلانياً لأن تنقيب البيانات يستخدم أيضاً الأفكار والأدوات والتقنيات من أجزاء أخرى ، لا سيما تكنولوجيا قواعد البيانات

والتعلم الآلي ، ولا يرتبط ارتباطًا كبيرًا ببعض الأجزاء التي يهتم بها الإحصائي والخوارزميات الإحصائية الأساسية هي التقنيات الوصفية والتصور ، تحليل الكتلة ، تحليل الارتباط ، تحليل التمييز ، تحليل العوامل ، تحليل الانحدار ، الانحدار اللوجستي.

8.2 مهام وأساليب تنقيب البيانات

في الواقع ، فإن الهدفين الرئيسيين لاستخراج البيانات هما "التنبؤ" و "الوصف يشارك التوقع في تطبيق العديد من الحقول أو المتغيرات ضمن مجموعة البيانات من أجل التنبؤ بقيم مجهولة أو مستقبلية للمتغيرات الأخرى المطلوبة. ومع ذلك ، يركز الوصف على اكتشاف أنماط تكشف البيانات القابلة للتفسير.

تختلف أهمية المقارنة للتنبؤ والوصف لتطبيقات استخراج البيانات الدقيقة بشكل ملحوظ بشكل عام ، يمكن التعبير عن العديد من المشكلات في المجالات المنطقية والأكاديمية والاقتصادية والتجارية من حيث المهام .

- **تصنيف**

يشمل التصنيف استكشاف خصائص الغرض المشار إليه مؤخرًا ونقله إلى واحدة من مجموعة محددة من الفئات، يوصف التصنيف بتفسير دقيق للفصول ، ومجموعة تدريب تغطي الحالات المصنفة مسبقًا، تقنيات التصنيف الرئيسية هي .

- **شجرة القرار:**

شجرة القرار هي أداة دعم قرار تستخدم رسمًا توضيحيًّا شبيها بالشجرة للقرارات والتبعات المتوقعة لها، متضمناً احتمال تحقق المخرجات، وكلفة الموارد، والمنفعة. هي رسم باتجاه واحد لعرض الخوارزمية. تستخدم شجرة القرارات عموماً في بحوث العمليات، خصوصاً في تحليل القرارات للمساعدة في تحديد الاستراتيجية التي ستؤدي لتحقيق الهدف. ويكيبيديا

- **الشبكات العصبية**

هي شبكة الخلايا العصبية، أو بالتوجه الحديث الشبكة العصبية الاصطناعية تتكون الشبكة العصبية الاصطناعية من خلايا عصبية أو عقد صناعية. الشبكة العصبية إما أن تكون شبكة عصبية بيولوجية مصنوعة من أعصاب بيولوجية أو أن تكون شبكة عصبية صناعية لحل مشاكل وقضايا الذكاء الاصطناعي.

- **تحليل الارتباط**

اكتشاف العلاقات المثيرة للاهتمام بين المتغيرين في قواعد البيانات الكبيرة. الغرض منه هو تحديد القواعد القوية المكتشفة في قواعد البيانات باستخدام بعض المقاييس، ويولد ذلك قواعد جديده لأنه يحلل المزيد من البيانات الهدف الاساسي فهم قدرات الارتباط التجريدي من البيانات الجديدة غير المصنفة بافتراض وجود مجموعه بيانات كبيره بما فيه الكفاية.

- **أقرب تقنيات الجار**

تعتبر خوارزمية الجار الأقرب (Nearest Neighbor Algorithm) من تقنيات التنقيب في البيانات، وهي من خوارزميات التصنيف والتنبؤ التي تهدف للتنبؤ عن طريق مقارنة السجلات الشبيهة بالسجل المراد التنبؤ له وتقدير القيمة المجهولة لهذا السجل بناء على معلومات لتلك السجلات.

إن أحد الأمثلة الواقعة لخوارزمية الجار الأقرب هي انه لو نظرت الى جيرانك الذين يسكنون حولك تلاحظ ان لديهم دخل متشابه، لذت فإنه لو كان جارك لديه دخل (10000) سنوياً فإن احتمال ان يكون دخلك انت أيضا (10000) سنوياً هو احتمال كبير، وربما الاحتمال كبير جدا ان دخلك بهذا الشكل إذا ما كان كل جيرانك كذلك، في حين يكون احتمال أقل بكثير عندما يكون دخلهم لا يتعدى (5000) فقط،

إذن بهذه الطريقة أمكننا تخمين مقدار دخل شخص ما بمجرد النظر لجيرانه القريبين منه.

- **دعم ناقلات الآلات**

دعم ناقل الآلات (SVM) هو خوارزمية تعلم آلة تحت إشراف والتي يمكن استخدامها لكليهما تحديات التصنيف والانحدار .ومع ذلك، يستخدم في الغالب في مشاكل التصنيف. في هذه الخوارزمية، نرسم كل عنصر من عناصر البيانات كنقطة في مساحة الأبعاد) n حيث يكون n هو عدد الميزات التي لديك (وتكون قيمة كل ميزة هي قيمة إحداثيات معينة. بعد ذلك، نقوم بإجراء التصنيف من خلال إيجاد الطاقة الفائقة التي تميز الفصلين.

2.8.2 تقدير

تتواءم التقديرات مع النتائج ذات القيمة المستمرة، يشبه تقدير التقدير التصنيف باستثناء أن المتغير المستهدف عدديًا وليس قاطعيا يتم استخدام التقدير لتعيين قيمة للعديد من المتغيرات المستمرة غير المحددة مثل الدخل أو الارتفاع أو رصيد الائتمان مهام التقدير المناسبة هي:

1. نماذج الانحدار
2. الشبكات العصبية
3. تحليل البقاء على قيد الحياة

3.8.2 التنبؤ

التنبؤ مطابق للتصنيف أو التقدير ، إلا أنه بالنسبة للتنبؤ ، فإن النتائج تكمن في المستقبل. في مهمة التنبؤ ، فإن الطريقة الوحيدة للتحكم في دقة التصنيف هي البقاء والمراقبة السبب الرئيسي لمعالجة التنبؤ كمهمة مختلفة عن التصنيف والتقدير هو أنه في النمذجة التنبؤية هناك مواضيع إضافية تتعلق بالعلاقة الزمنية لمتغيرات المدخلات أو المتنبئين بالمتغير المستهدف، يمكن تخصيص أي من الطرق المستخدمة

للتصنيف والتقدير للتطبيق في التنبؤ عن طريق حالات التدريب التي يتم فيها تحديد قيمة المتغير الذي سيتم التنبؤ به ، بالإضافة إلى البيانات التاريخية لتلك الحالات، يتم استخدام البيانات التاريخية لإنشاء نموذج يوضح السلوك الحالي الملاحظ. بينما يتم استخدام هذا النموذج لتقديم المدخلات ، فإن النتيجة هي توقع لسلوك المستقبل بعض الأمثلة على مهام التنبؤ في مجال الأعمال والبحوث هي:

توقع المبيعات المستقبلية للشركة ؛

توقع العملاء الذين سوف يختفون في الأشهر المقبلة ؛

التنبؤ بمشتركي الهاتف الخليوي الذين قد يرغبون في طلب خدمة ذات قيمة مضافة في المستقبل.

بعض تقنيات التنبؤ الشهيرة هي:

1. الانحدار الخطي
2. الانحدار غير الخطي
3. شجرة القرار
4. طرق التنبؤ التقليدية مثل ARMA أو ARIMA
5. الشبكات العصبية

عادةً ، "يمكن أيضاً استخدام أي من الأساليب والتقنيات المستخدمة للتصنيف والتقدير للتنبؤ ، في ظل الظروف المناسبة ، للتنبؤ". يعتمد اختيار الطريقة على طبيعة بيانات المدخلات ، ونوع القيمة التي يجب التنبؤ بها ، والأهمية المرتبطة بتفسير التنبؤ، سنستخدم الشبكات العصبية للتنبؤ بالمبيعات المستقبلية نظراً لقدرتها على تخزين واستخدام المعلومات حول البيانات السابقة وإجراء تنبؤات دقيقة. ستتم مناقشة وصف وتطبيقات الشبكات العصبية في التنبؤ لاحقًا.

- **أنماط التنقيب المتتابعة**

ان استكشاف الانماط المتكررة بكل انواعها يلعب دورا مهما في تنقيب واستكشاف علاقات التبعية والارتباطات والعلاقات الاخرى الشيقة بداخل قاعدة البيانات، وباستخدام تقنيات استكشاف الانماط تستطيع المؤسسات والشركات بكافة انواعها استكشاف العلاقات والارتباطات الخفية الشيقة بداخل قواعد البيانات وتبسيطها وتوضيحها لصناع القرار من اجل مساعدتهم في تحسين الادارة واتخاذ القرارات ورفع كفاءة التخطيط وتحسين خطط التسويق المتعدد الاغراض وتحليل وادارة وتقييم المخاطر بطرق البيانات واجراءات التنقيب الاخرى للكميات الهائلة من البيانات التي يتم تجميعها وتخزينها بصفة مستمرة.

وقد اصبحت عمليات تنقيب الانماط المتكررة من العمليات المهمة في تنقيب البيانات ويتم التركيز عليها كأحد مسارات بحوث تنقيب البيانات بشكل عام.

الانماط المتتابعة يمكن ان تكون أحد مما يلى:-

- مجموعة من العناصر
- او الفئات المتسلسلة
- او البناءات الجزئية

- **البحث على أساس تشابه البحث**

في معظم تقنيات تحليل وتنقيب البيانات، مثل خوارزمية التحليل العنقودي وخوارزمية الجار الاقرب تظهر الحاجه الى قياس تشابهه واختلاف البيانات من اجل تقييم مدى التشابه والاختلاف فيما بين البيانات، مثلا قد يحتاج احد المراكز الى تجزئة زبائنه الى مجموعات ذات خصائص مميزة، كأن يقوم بتجميع الزبائن المتشابهين في الدخل او العمر بحيث يمكن استخدام هذه المعلومات في وضع خطط واستراتيجيات التسويق المختلفة التي تستهدف فئات وشرائح معينة من الجمهور.

مثلا في احد الخوارزمية التي تستكشف انواع الامراض وتشخيصها قد يتم اللجوء لبحث التشابه في الاعراض التي تظهر على مرضى اخرين مصابين بالفعل بمرض معين، وبالتالي التوصل الى التشخيص الصحيح لحالة المريض.

- **التغيير وكشف الانحراف**

مهمة استخراج البيانات التي يكون الهدف منها هو بناء نموذج يصف اهم التغييرات في البيانات من القيم المعيارية التي يتم قياسها مسبقا.

- **اختيار تقنيات البيانات المناسبة**

في الابحاث التي تستهدف مجتمعات عدد افرادها كبير، يصعب استخلاص المعلومات والبيانات من كل افراد المجتمع، لذلك يتم اختيار عينة تمثيلية عن المجتمع. ولاختيار العينة المناسبة للبحث لابد من الانتباه الى مجموعة من الخصائص المطلوبة، يأتي على راسها العينة حيث يمكن من خلالها تمثيل هذا المجتمع وحمل نفس خصائصه، وتعرف العينة بأنها مجموعة من الأشخاص الذين يحددهم الباحث، للمساهمة في البحث، وليست مجموعة الأشخاص الذين يستوجب عليهم المشاركة في البحث، ويقع تحديد العينة المناسبة للبحث ضمن شروط متعددة منها توفر صفات وخصائص المجتمع الأصلي في العينة. ولا يوجد توافق بين عدد أفراد العينة وعدد الأفراد الذين يتضمنون في المجتمع الأصلي. وكذلك منح جميع أفراد المجتمع الأصلي فرصة متكافئة لأن يتم انتقاؤهم للانضمام لعينة البحث.

خطوات اختيار العينة البحثية المناسبة

اختيار المجتمع الأصلي

والذى نحدد منه العينة، فالمجتمع هو الهدف الأساسي من البحث، حيث يعمل الباحث على تعميم النتائج عليه في النهاية، ويمكننا القول بأننا لا ندرس عينات وإنما ندرس مجتمعات. وما العينة التي نحددها إلا وسيلة لتحليل صفات المجتمع بأكمله. ولذلك فإن الخطوة الرئيسية في تحديد العينة هي التعرف على خصائص وصفات المجتمع

. ويشمل تعريف المجتمع صفة واحدة على الأقل تميزه عن غيره من المجتمعات. والهدف من تعريف المجتمع هو تحديد عدد ما يتضمنه من أشخاص.

التعرف على صفات المجتمع

عند تحديد خصائص المجتمع نحدد قائمة بهذه الصفات من وجهة نظر البحث، أي من وجهة نظر المتغيرات التي تتضمنها الدراســة مثل (العمر– الجنس – المنطقة التعليمية – الحالة الاجتماعية – المهنة .(وهذه الخصائص هي فقط للأشخاص، وهناك مجتمعات اخرى مثل مجتمعات نوع من النباتات أو المركبات الكيميائية أو المادية لها خصائص معينة. فعلى سبيل المثال؛ عند التعرف على نوع التربة المناسب لزراعة معينة فتكون للتربة خصائص أخرى مثل اللون والخصوبة ودرجة الامتصاص وغيرها. ومن الطبيعي أن تختلف هذه الخصائص وفقاً لغايات وأهداف الدراسة.

تحديد حجم المجتمع

يوجد عدة قوانين متعلقة بتحديد حجم العينة، حيث تمتلك كل دراسة غاياتها وأهدافها، ولكن ينصح الإحصاء الاستدلالي بزيادة حجم العينة، حيث ترتفع فرص التمثيل عند زيادة حجم العينة، ويجد الباحث نفسه أمام اختيارين:

إما أن تكون العينة صغيرة نسبيا يكون التعامل معها سهل من جميع الجوانب " تحديد المتغيرات – قلة نسبة التكاليف – سرعة التوصل إلى النتائج. لكن في هذه الحالة العينة لا تمثل المجتمع ولا جدوى من الدراسة).

أو أن يحدد عينة كبيرة ذات فرص تمثيل مرتفعة، لكن يصعب التحكم بالمتغيرات لكبر عددها، ولتفاعلها مع بعضها البعض بشكل قد لا يمكن توقعه مسبقا، فضلاً عما يتحمله الباحث من جهد ووقت ونفقات.

ويتوقف حجم العينة المناسبة للبحث على عدة عوامل في البحث مثل نوع المجتمع الأصلي، ونوع البحث، وفروض البحث، وتكاليف البحث، وأهمية النتائج، وطرق

جمع البيانات والدقة المطلوبة في البحث. وعلى كل الأحوال فإن اختيار العينة المناسبة للبحث يحتاج إلى بعض الخبرة لحصر هذه العوامل كلها.

- **تطبيقات استخراج البيانات**

استخراج البيانات هو عملية الحصول على البيانات من مصدر لمزيد من معالجة البيانات او تخزينها او تحليلها في مكان اخر ومثال لتطبيقات استخراج البيانات Orange، Rapidminer،Weka

- **بيانات الشبكة العصيبوتية**

قد ساهمت طبيعة خوارزميات الشبكات العصبية في ان تكون هي الاكثر استخداما في مجال الذكاء الاصطناعي باعتبار انها تهدف الى محاكاة الذكاء البشرى واكساب الالة بعض قدرات العقل الطبيعي. ويتم بناء خوارزميات الشبكات العصبية بطرق معقدة جدا ومبنية على انواع البيانات الرقمية، وعادة ما يتم استبدال قيم المتغيرات الاسمية المتوفرة بقيم رقمية لكى يتم استخدامها في الشبكة العصبية، وذلك من خلال عمليات تحويل البيانات وتفريد البيانات التي تتم في مرحلة تحضير البيانات للتحليل والتنقيب.

ومع ذلك ، فإن المهام الأساسية التي يمكن القيام بها مع استخراج البيانات هي: التصنيف ، التقدير ، التنبؤ ، قواعد التجميع أو الارتباط ، التجميع ، الوصف والتصور، يمكننا أن نرى شرحا موجزا للوظائف المذكورة أدناه.

النماذج المختارة في هذا البحث (شجرة القرار).

قواعد التقارب أو التجمع

تتمثل مهمة قاعدة الاتحاد في استنتاج العناصر التي تسير معًا، يتم استخدامها عادةً في مجتمع مبيعات التجزئة لتحديد الأشياء التي يتم شراؤها عادة معًا تتمثل مهمة الارتباط في اكتشاف قواعد لتحديد العلاقة بين ميزتين أو أكثر"تتخذ قواعد الشراكة النموذج إذا كانت سابقة ، فتبعًا لذلك ، إلى جانب قدر من الدعم والثقة المرتبط

بالقاعدة". يمكن أيضًا تطبيق قواعد الشراكة للتعرف على آفاق البيع المتقاطع ولتخطيط الحزم المذهلة أو الجمع بين المنتجات والخدمات بعض خوارزميات قاعدة الارتباط هي:

الانحدار الخطي

1. HPTree
2. AIS
3. SETM
4. RARM

التجميع (التعلم غير الخاضع للإشراف):

"التجميع هو مهمة تقسيم السكان غير المتجانسين إلى عدد من المجموعات الفرعية أو المجموعات المتجانسة يشير "التجميع" إلى تجميع السجلات أو الملاحظات أو الحالات في فئات كائنات متشابهة ".

الكتلة هي مجموعة من السجلات التي تشبه بعضها البعض ، وتختلف عن السجلات في الكتل الأخرى في الواقع ، ما يميز المجموعات عن التصنيف هو أن التجميع لا يعتمد على فئات محددة مسبقًا ومع ذلك ، في التصنيف يتم تخصيص كل سجل لفصل محدد مسبقًا على أساس نموذج تم تطويره أثناء التدريب على الحالات المصنفة مسبقًا" يتم تنفيذ التجميع غالبًا كخطوة أولية في عملية استخراج البيانات ، مع استخدام المجموعات الناتجة كمدخلات إضافية في تقنية مختلفة في اتجاه المصب، مثل الشبكات العصبية.

الوصف والتنميط:

في بعض الأحيان ، يكون هدف استخراج البيانات هو توضيح ما يجري في قاعدة بيانات معقدة من خلال نهج يضيف إلى فهمنا للأشخاص أو المنتجات أو الإجراءات التي خلقت البيانات في المكان الأول حيث يمكن تحقيق الوصف المتميز عن طريق

تحليل البيانات الاستكشافية، وهي تقنية رسومية لاكتشاف البيانات في البحث عن الأنماط والاتجاهات.

9.2 اختيار نموذج التنقيب عن البيانات المناسب:

شجرة القرار :

شجرة القرار عبارة عن تمثيل بياني لعملية القرار وتتكون هذه الشجرة من العناصر التالية:

نقاط القرار ، البدائل ، نقاط الفرص أو الحدث ، حالات الطبيعة ، والعوائد.

تعتبر شجرة القرارات من الأدوات التي يعتمد عليها متخذي القرار في حل المشكلات ، خاصة في حالة أن يمر حل المشكلة بعدة مراحل ، كما أن شجرة القرارات تساعد على استخدام الاحتمالات المشتركة واللاحقة للتوصل إلى أفضل حل للمشكلة ، إن شجرة القرارات تبدأ دائماً بنقطة قرار ، والتي تمثل في النهاية القرار الذي سوف نتوصل له لحل المشكلة.

ويوجد في شجرة القرار نوعين من المنابت مربع يمثل نقطة قرار ودائرة تعبر عن حدث صدفة (أي عشوائي) ويجب أن تشتمل بيانات شجرة القرار على الاحتمالات الخاصة بالفروع التي تخرج من منابت الأحداث والإيرادات الخاصة بالبدائل المختلفة للمشكلة

التنبؤ باستخدام الانحدار الخطى

الانحدار أو يسمى التنبؤ Prediction وهو تقدير القيمة المستقبلية لمتغير واحد بناءً على معرفة قيم متغير أو أكثر، وهنالك عدة أنواع من معامل الانحدار:

1) الانحدار الخطي Linear Regression تشير تسمية هذا المعامل الى أنه يتضمن متغير تابع y يعتمد على متغير واحد مستقل x وكلمة خطي تشير الى أن العلاقة ين المتغيرين y و x هي علاقة خطية.

2) الانحدار المتعدد Multiple Linear Regression هذا النوع من الانحدار يتضمن اعتماد المتغير y على أكثر من متغير مستقل مثل x1 و x2 ... الخ.

3) الانحدار غير الخطي Non-Linear Regression إذا كانت العلاقة بين المتغير y والمتغيرات المستقلة غير خطية مثل علاقة أسية أو لوغاريتمية أو تربيعية ... الخ. وهنالك أنواع أخرى مثل الانحدار الهرمي Hierarchical Regressionوالانحدار التدريجي Stepwise Regression وغيرها.

الانحدار الخطي Linear Regression

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

الفصل الثالث

منهجية البحث

الفصل الثالث

منهجية البحث

في هذا الفصل ، يتم شرح منهجية البحث المستخدمة لتحقيق الأهداف والإجابة على أسئلة البحث. يبدأ بتصميم البحث. يوضح الغرض من البحث ، نهج البحث ، واستراتيجية البحث. ثم، تليها عملية البحث التي استخدمت في هذا البحث. جميع الطرق المطبقة في كل مرحلة من مراحل النمذجة: جمع البيانات ، ومعالجة البيانات، والحصول على نتائج التنبؤ بالنماذج المستخدمة في هذا البحث وتقييم النماذج المذكورة.

1.3 غاية البحث

يمكن تصنيف الأبحاث حسب الغرض منها. على سبيل المثال ، تصنف الأبحاث في أغلب الأحيان على أنها تحليلية ونهائية. يمكن تصنيف التصميمات البحثية الشاملة إلى التحليلية (التوضيحية) والوصفية. الأنواع المختلفة موضحة أدناه: كما يشير اسمها ، تهدف البحوث التحليلية إلى استكشاف مشكلة أو موقف أو البحث فيه من أجل توفير رؤية وفهم. بمعنى آخر ، "الغرض الأساسي من البحث التحليلي هو تسليط الضوء على طبيعة الموقف وتحديد أي أهداف أو بيانات محددة تحتاج إلى معالجة من خلال بحث إضافي". "البحوث التحليلية مفيدة للغاية عندما يرغب صانع القرار في فهم موقف ما و / أو تحديد بدائل القرار بشكل أفضل. يكون التحليل مفيداً بشكل خاص عندما يفتقر الباحثون إلى فكرة واضحة عن المشكلات التي سيواجهونها أثناء الدراسة ". ومع ذلك ، فإن الغرض من البحث الوصفي هو شرح خصائص. يمكن إجراء البحث التحليلي للتنبؤ أو لمعرفة الدرجة التي ترتبط بها المتغيرات. "الوصف هو جعل الأشياء المعقدة مفهومة عن طريق اختزالها إلى الأجزاء المكونة لها. يمكن أن يكون البحث الوصفي على صلة مباشرة بالبحث التحليلي ، لأن الباحثين ربما يكونوا قد بدأوا بالرغبة في اكتساب نظرة ثاقبة لمشكلة

وبعد ذكرها ؛ أصبحت أبحاثهم وصفية ". أخيراً ، أقامت الدراسات السببية العلاقة السببية بين المتغيرات. في هذه الدراسات ، يتم التركيز على تحليل الموقف أو المعضلة لتوضيح العلاقات بين المتغيرات.

في الواقع ، يفسر تعبير "استخراج البيانات" إجراء اكتشاف المعرفة أو المعلومات من قواعد البيانات المخزنة في مستودعات البيانات. الغرض من التنقيب عن البيانات هو اكتشاف واستخراج أنماط أو نماذج صالحة وجديدة وذات مغزى في البيانات. يعد استخراج البيانات أداة قيّمة ، وهو نهج يدمج الاستكشاف والاكتشاف مع دراسة إحصائية مؤكدة لاكتشاف العلاقات والمصادقة عليها. ومع ذلك ، يمكننا تعيين مهام استخراج البيانات في واحدة من فئتين.

أولاً ، تنبأ بيانات التنقيب عن ذلك

"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 اكتشاف الأنماط التي يمكن تفسيرها من قبل الإنسان وتحديد الخصائص العامة للبيانات في قاعدة البيانات. قد تساعد المهام الوصفية أيضًا الأبحاث في التنبؤ بها (مثل بحثنا). سنرى في الفصل التالي قيامنا بالمهام الوصفية والتنبؤية. نظراً لأن أداتنا في هذه الدراسة هي التنقيب عن البيانات ، وسوف نقوم بتطبيق كل من المهام الوصفية والتنبؤية لاستخراج البيانات ، فإن الغرض من هذا البحث هو تحليلي بشكل أساسي ولكن مع بعض الجوانب الوصفية التي تساعدنا في مرحلة التنبؤ الخاصة بنا.

وتعمل منهجية البحث على تحقيق أغراض البحث عن طريق أساليب متنوعة: النهج الكمية والنوعية. يقدم البحث النوعي رؤى وإدراك لمشكلة البحث. خصائص الدراسة النوعية هي أنها تستند إلى حد كبير على وصف الباحث الخاص ، والعواطف وردود الفعل. من ناحية أخرى ، يسعى البحث الوصفي إلى تحديد البيانات. تسعى للحصول

على أدلة قاطعة من خلال تطبيق بعض أشكال التحليل الإحصائي. وهذا يعني ، في النهج الوصفي ، أن النتائج تستند إلى أرقام وإحصائيات معروضة في أرقام ، بينما في المناهج النوعية تستند النتائج إلى شرح حدث بتطبيق الكلمات. بالإضافة إلى ذلك ، يكمن البحث النوعي في النهاية غير المنظمة للاستمرارية ، في حين أن البحث الكمي منظم للغاية.

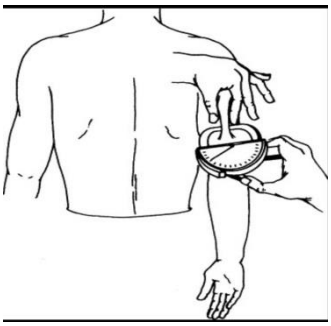
على الرغم من حقيقة أن الاستكشاف يعتمد بدرجة كبيرة على التقنيات النوعية، إلا أنه في العديد من الحالات توجد بيانات كافية للسماح باستخراج البيانات أو استكشاف العلاقات بين القياسات الفردية. يتيح مفهوم استخراج البيانات دعم صناع القرار عن طريق البحث الوصفي التحليلي. نظراً لأن منهج هذه الدراسة هو استخراج البيانات ، يمكن اعتبار هذا البحث بمثابة بحث وصف

2.3 تحليل البحث

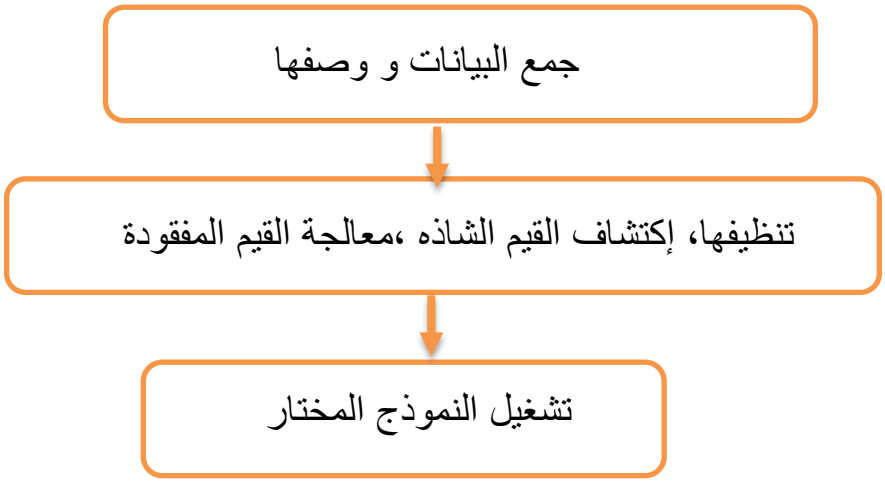
في هذا البحث سوف نقوم بجمع البيانات اللازمة للقيام بالتنبؤ بمرض السكري. بناءً على الدراسات السابقة فإن البيانات المطلوبة هي :

Name	Description
Pregnancies	No of Times Pregnant
Glucose	Plasma glucose concentration
blood pressure	Diastolic Blood Pressure
skin thickness	Triceps skin folds thickness
Insulin	2-Hours Serum Insulin
BMI	Body Mass Index
Pedigree Diabetes Function	Diabets Pedigree Function
Age	Age
Outcome	Diabetes Mellitus Type II

skin thickness

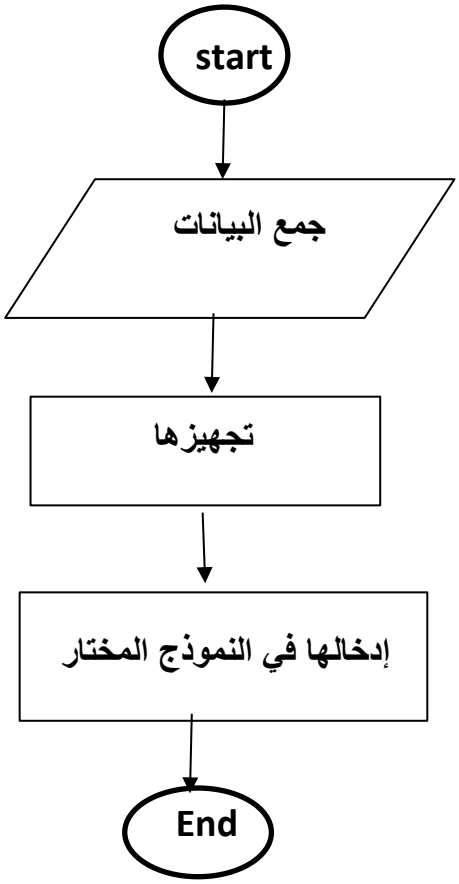


بعد الحصول عليها و جمعا سنقوم بي عملية معالجة لها لتكون صالحة للإدخال في النماذج المختارة (شجرة القرار و الانحدار الخطي)



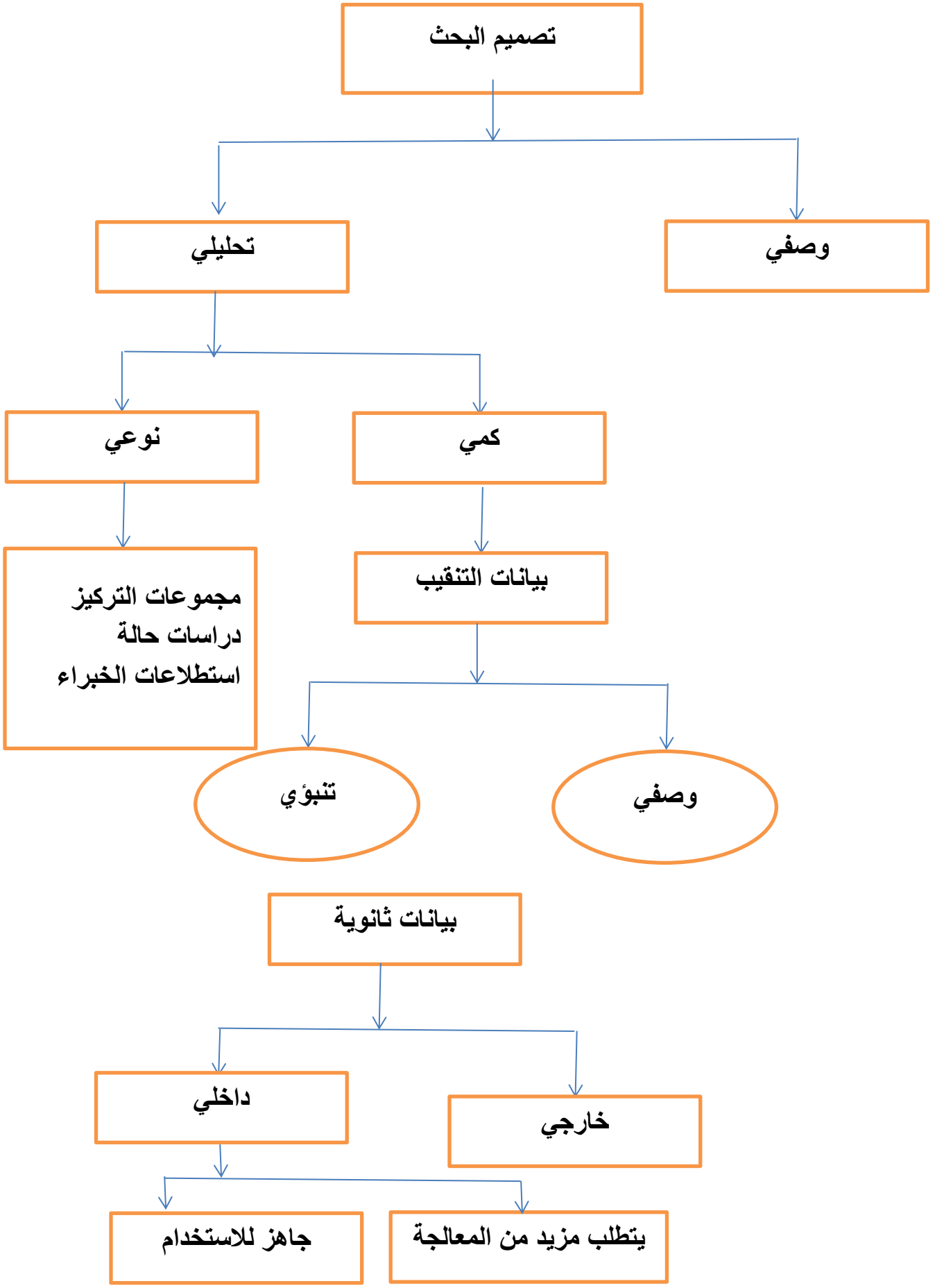
3.3 المخطط الانسيابي :

هو رسم توضيحي لخطوات معينة. نشأت من علوم الحاسوب كأداة لتمثيل الخوارزميات و منطق البرمجة و لكنها امتدت لاستخدامها في جميع أنواع العمليات الأخرى.



4.3 تصميم البحث

تصميم البحث هو إطار لإنجاز البحوث وهو الأساس لتنفيذ المشروع. وفقًا لذلك ، فهي خطة أساسية توجه أجزاء جمع البيانات وتحليلها من البحث. وهي تحدد البيانات حسب نوع المعلومات التي سيتم جمعها ، ومصادر البيانات ، وعملية جمع البيانات "يتضمن التصميم البحثي الجيد أن تكون المعلومات التي يتم جمعها متوافقة مع أهداف الدراسة وأن الإجراءات المتعلقة بجمع البيانات دقيقة وفعالة.


تصميم البحث
تحليلي
وصفي
نوعي
كمي
مجموعات التركيز
دراسات حالة
استطلاعات الخبراء
بيانات التنقيب
تنبؤي
وصفي
بيانات ثانوية
داخلي
خارجي
جاهز للاستخدام
يتطلب مزيد من المعالجة

الفصل الرابع

التحليل و التصميم

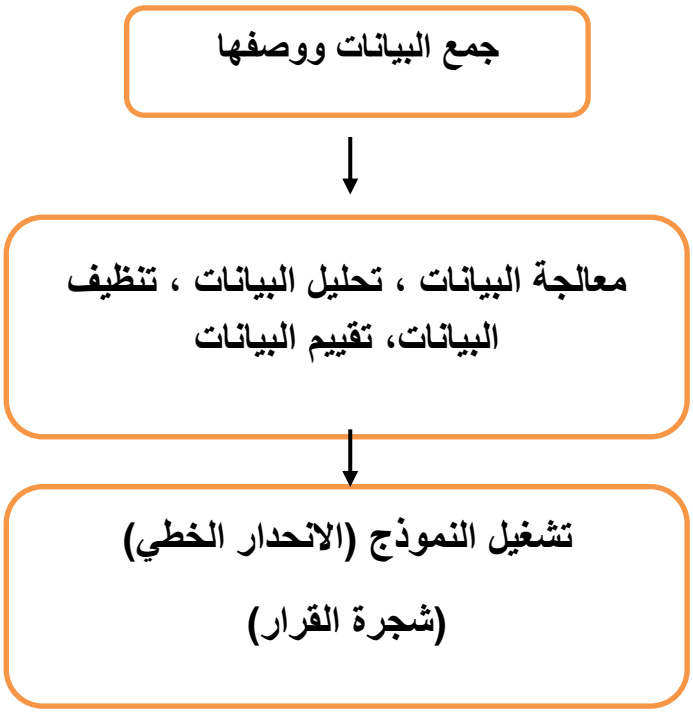
الفصل الرابع

التحليل و التصميم

يشرح هذا الفصل جمع البيانات ووصفها واعدادها وتحليلها، البيانات تم جمعها من مصدر عبر الانترنت(https://data.world/data-society/pima-indians-diabetes-database). والهدف من هذا البحث 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 الانحدار الحظي وشجرة القرار.

لأن نموذج الانحدار الخطي وشجرة القرار لديهم القدرة على التنبؤ.

في هذا البحث نستخدم أداة Rapid miner،" مصدر مفتوح برنامج الترميز ويعطي تحليلات متقدمة وسهلة الاستخدام في عملية استخراج البيات".



1.4 اعداد البيانات:

بعد جمع البيانات اللازمة يجب دمج البيانات وتنظيمها وتحويلها لتكون مناسبة للتنبؤ بمرض السكري.

لأن قواعد البيانات حساسة للغاية للبيانات الصاخبة والمفقودة وغير المنسقة، هنالك عدد من تقنيات معالجة البيانات الأولية، تنظيف البيانات، تكامل البيانات، تحويل البيانات، والحد من البيانات، يمكن تطبيق تنظيف البيانات لإزالة الشواذ، والقيم المفقودة في البيانات.

تدمج البيانات من مصادر متعددة في مخزن بيانات واضح، البيانات تتضمن التحويلات تطبيع/ قياس وبنية المعالم. يتم تحويل البيانات وتوحيدها في هياكل مناسبة للتنقيب.

يمكن تقليل حجم البيانات بواسطة الجمع بين إزالة الميزات الزائدة (اختيار الميزة)، أو التجميع، يمكن ان تعمل هذه الطرق بشكل متبادل، عندما تستخدم قبل التنقيب يمكن ان تحسن إلى حد كبير، الجودة العامة للانحدار أو الوقت اللازم للتنقيب الحقيقي، تستغرق خطوة معالجة البيانات عادة معظم الوقت.

Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1

2.4 تنظيف البيانات:

في هذه العملية، فإن القيم المفقودة والمكررة والبيانات غير المفيدة سوف يقوم بمعالجتها عمليات Rapid miner المستخدمة هي (استبدال القيم المفقودة، ازالة التكرار) في هذه البيانات لم يكن هنالك قيم مفقودة ولا توجد قيم مكررة.

كان ملف Excel يحتوي على السمات المطلوبة.

3.4 اكتشاف القيمة الشاذة:

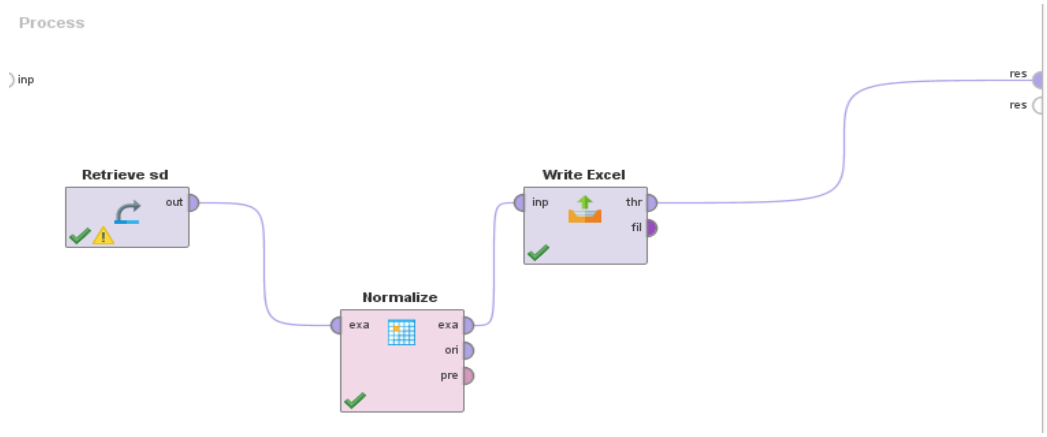
القيمة الشاذة (outlier) هي عنصر شاذ وخارج عن النسق المميز لمجموعة او تركيب معينة.

الأدوار المطبقة في البيانات هي:

1. تخزين البيانات.
2. تحويل البيانات.
3. تطبيع مقياس البيانات.

تطبيع البيانات منطوي على توسيع نطاق قيم السمات لجعلها عدديا وبالتالي يكون له نفس الأهمية، شجرة القرار تتيح نماذج أفضل عندما يتم تطبيع البيانات وسيساعد تطبيع بيانات الادخال في تسريع مرحلة التدريب ويجب ان تكون جميع البيانات موحدة قبل النمذجة.

1. استرداد البيانات التي تم تنظيمها.
2. تطبيع



أ. **نوع مرشح (السمة) (مجموعة فرعية):**

تسمح لك هذه المعلمة بتحديد مرشح تحديد سمات، لديها خيارات التالية:

- **الكل:** يحدد هذا الخيار جميع سمات ال Example بحيث لا تتم إزالة السمات وهذا هو الخيار الافتراضي.
- **مفرد:** يسمح هذا الخيار بتحديد سمة واحدة، المطلوب يتم تحديد السمة بواسطة النوع يتيح هذا الخيار اختيار sub set سمات متعددة من خلال قائمة ، إذا كانت بيانات التنقيب (Example set) معروفة لجميع السمات الموجودة في القائمة ويمكن ان تكون الخصائص المطلوبة بسهولة المحدد.

السمة المحددة في هذا البحث هي (الكل).

ب. **الطريقة:**

طريقة (اختياري)

يتم توفير أربع طرق هنا لتطبيع البيانات. هذه الطرق مو ضحة أي ضًا في العملية التعليمية المرفقة.

z_transformation: يسمى هذا أيضًا التطبيع الإحصائي. يطرح هذا التسوية متوســـط البيانات من جميع القيم ثم يقســـمها على الانحراف المعياري. بعد ذلك ، يكون لتوزيع البيانات متوسط صفر ومتباين بواحد. هذه تقنية تطبيع شائعة ومفيدة للغاية. يحافظ على التوزيع الأصلي للبيانات وأقل تأثرًا بالقيم المتطرفة.

range_transformation: يؤدي تحويل النطاق إلى تسوية جميع قيم السمات إلى نطاق قيمة محدد. عند تحديد هذه الطريقة ، تظهر معلمتان أخريان (min، max) في لوحة Parameters. لذلك يتم تعيين أكبر قيمة على "max" ويتم تعيين أصغر قيمة على "min". يتم قياس جميع القيم الأخرى ، بحيث تتناسب مع النطاق المحدد. يمكن أن تتأثر هذه الطريقة بالقيم المتطرفة ، لأن الحدود تتحرك نحوها. من ناحية أخرى ، تحافظ هذه الطريقة على التوزيع الأصـــلي لنقاط البيانات ، بحيث يمكن استخدامها أيضًا لإخفاء هوية البيانات ، على سبيل المثال لإخفاء النطاق الحقيقي للملاحظات.

نسبة_التحول: يعتمد هذا التسوية على نسبة كل قيمة سمة على السمة الكاملة. هذا يعني أن كل قيمة مقسومة على المجموع الإجمالي لقيم هذه السمة. يتكون المجموع فقط من القيم المحدودة ، متجاهلاً القيم NaN / المفقودة وكذلك اللانهاية الموجبة والسالبة. عند تحديد هذه الطريقة ، تظهر معلمة أخرى (السماح بالقيم السالبة) في لوحة Parameters. إذا تم تحديده ، فسيتم التعامل مع القيم السالبة كقيم مطلقة ، وإلا فإنها ستنتج خطأ عند تنفيذها.

interquartile_range: يتم إجراء التطبيع باستخدام النطاق الربيعي. النطاق الربيعي هو المسافة بين المئين الخامس والعشرين والخامس والسبعين، والتي تسمى أيضًا الربيع الأدنى والأعلى، أو Q1 و Q3. يتم حسابها بفرز البيانات أولاً ثم أخذ قيمة البيانات التي تفصل بين أول (أو آخر) 25٪ من الأمثلة عن الباقي. الوسيط هو المئين الخمسين، لذا فهو القيمة التي تفصل القيم التي تم فرزها إلى النصف.

النطاق الربيعي (IQR) هو الفرق بين Q3 و Q1. الصيغة النهائية لتطبيع النطاق الرباعي هي: (متوسط القيمة) / معدل الذكاء IQR هو النطاق بين متوسط 50٪ من البيانات ، لذلك فإن طريقة التطبيع هذه أقل تأثراً بالقيم المتطرفة. سيتم تجاهل NaN / القيم المفقودة ، وكذلك القيم اللانهائية لهذه الطريقة. أيضًا ، إذا لم يتم العثور على قيم محددة ، فسيتم تجاهل السمة المقابلة

- **التحويل:**

هذا ما يسمى أيضا التطبيع الاخصائي، هذا التطبيع يطرح متوسط البيانات من جميع القيم ثم يقسم لهم من الانحراف المعياري، بعد ذلك توزيع البيانات لديها يعني صفر وتباين واحد هذا شائع ومفيد للغاية.

- **تقنية التطبيع:**

تحافظ على توزيع الأصل للبيانات واقل تأثيرا من القيم المتطرقة

4.4 أخذ عينات البيانات للتدريب والاختبار:

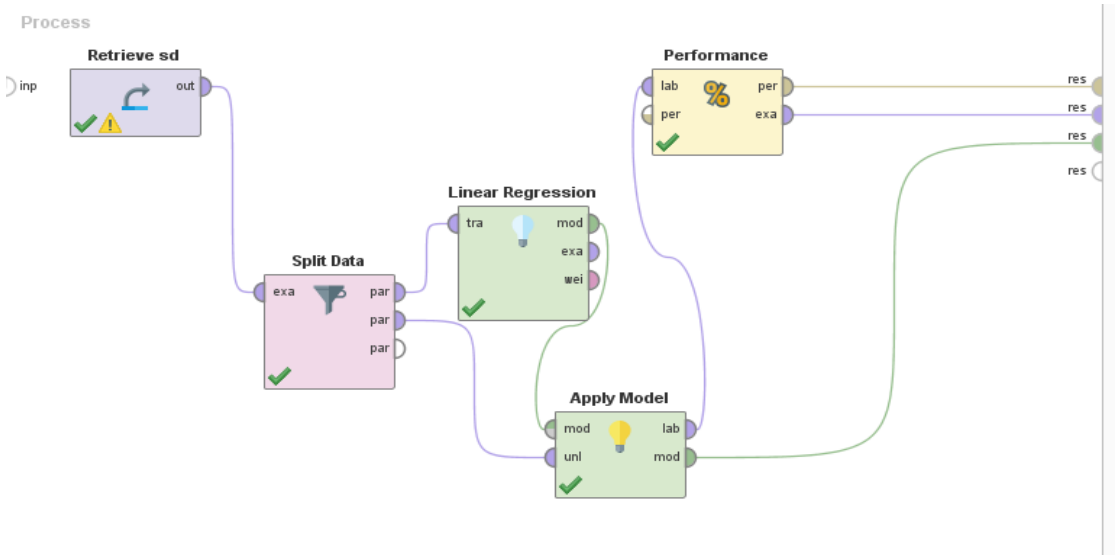
في هذه المرحلة يجب تقسيم البيانات إلى مجموعات تدريب واختبار في الواقع يتم تطبيق مجموعة لإنشاء النموذج، ويتم تطبيق مجموعة بيانات الاختبار لاختبار النموذج. للقيام بهذه المرحلة في Rapid miner يمكنني استخدام عامل تشغيل (تقسيم البيانات) وجعلها مقسمة إلى (3. & 7.) للتدريب و الإختبار.

5.4 الإنحدار الخطي

لتنفيذ الإنحدار الخطي داخل (Rapidminer) في Process interface نقوم بإدخال البيانات المعالجة (retrieve data) وتقسيمها إلى split) set training set and test data) ثم إدخالها إلى معامل الإنحدار الخطي (linear Regression) و معامل الـــ Apply model لتدريب النموذج على مجموعه من البيانات (training set) ثم توصـــيل نموذج الإنحدار الخطي بالـــ Apply model

و توصيل ال الـــ Apply model بي معامل الآداء (performance) للحصول على (accuracy،recall ، precision and F measure) الخا صة بالإنحدار الخطي كما هو موضح في الشكل ().

1. استرداد البيانات(DIABETES).
2. تقسيم البيانات (.3،.7).
3. عامل الانحدار الخطي.
4. تطبيق النموذج.
5. الأداء.



ConfusionMatrix:		
True:	1.0	0.0
1.0:	46	21
0.0:	34	129

Confusion matrix for linear regression

TN=46، TP=129، FN=21،FP=34

- الدقة(**accuracy**)=(TP+TN/TP+FP+FN+TN)=(129+46/46+21+34+129)=
76.09%
- **Precision**=(TP/TP+FP) =(129/129+34) = **79.14%**
- **Recall** = (TP/TP+FN) =(129/129+21) = **86.00%**
- **F measure** =(2TP/2TP+FP+FN) = **82.43%**

6.4 شرح الإنحدار الخطي:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

where:

Y_i = dependent variable

β_0 = population Y intercept

β_1 = population slope coefficient

X_i = independent variable

ε_I = Random Error term

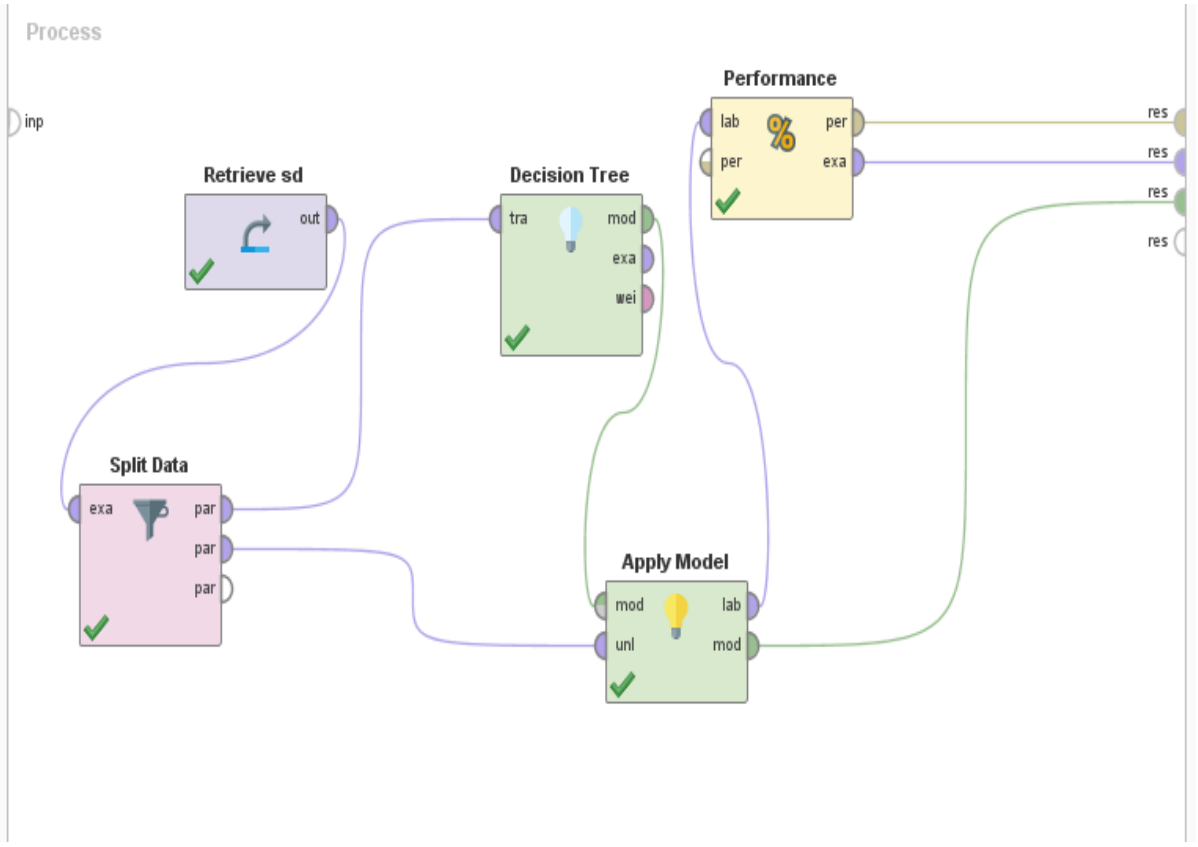
Attribute	Coefficient	Std. Error
Pregnancies	-0.019	0.006
Glucose	-0.006	0.001
BloodPressure	0.003	0.001
Insulin	0.000	0.000
BMI	-0.015	0.002
DiabetesPedigr...	-0.193	0.052
Age	-0.003	0.002
(Intercept)	1.897	0.098

```
- 0.019 * Pregnancies
- 0.006 * Glucose
+ 0.003 * BloodPressure
+ 0.000 * Insulin
- 0.015 * BMI
- 0.193 * DiabetesPedigreeFunction
- 0.003 * Age
+ 1.897
```

7.4 شجرة القرار

لتنفيذ شجرة القرار داخل (Rapidminer) في Process interfac نقوم بإدخال البيانات المعالجة (retrieve data) و تقسيمها إلى training set and test set (split data) ثم إدخالها إلى معامل شجرة القرار (decision tree) و معامل الـــ Apply model لتدريب النموذج على مجموعه من البيانات (training set) ثم توصيل نمذج شرة القرار بالـــ Apply model و توصيل ال الـــ Apply model بي معامل الآداء (performance) للحصول على(accurac، recall ، precision and Fmeasure) الخاصة بشجرة القرار

1. استرداد البيانات (بعد القاعدة).
2. تقسيم البيانات 7.)،(3.
3. عامل شجرة القرار.
4. تطبيق نموذج.
5. الأداء



Confusion matrix for decision tree

```
ConfusionMatrix:
True:    1.0      0.0
1.0:     24       11
0.0:     56       139
```

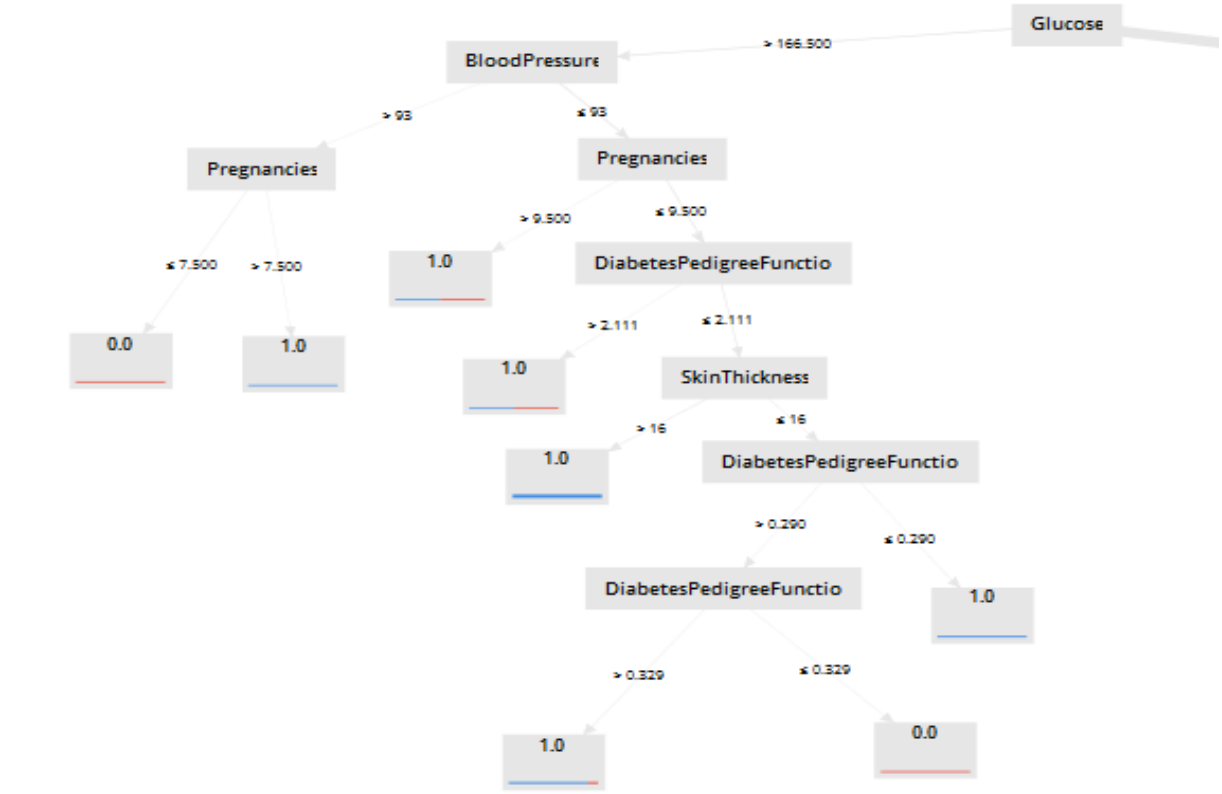
TN=24، TP=139، FN=11،FP=56

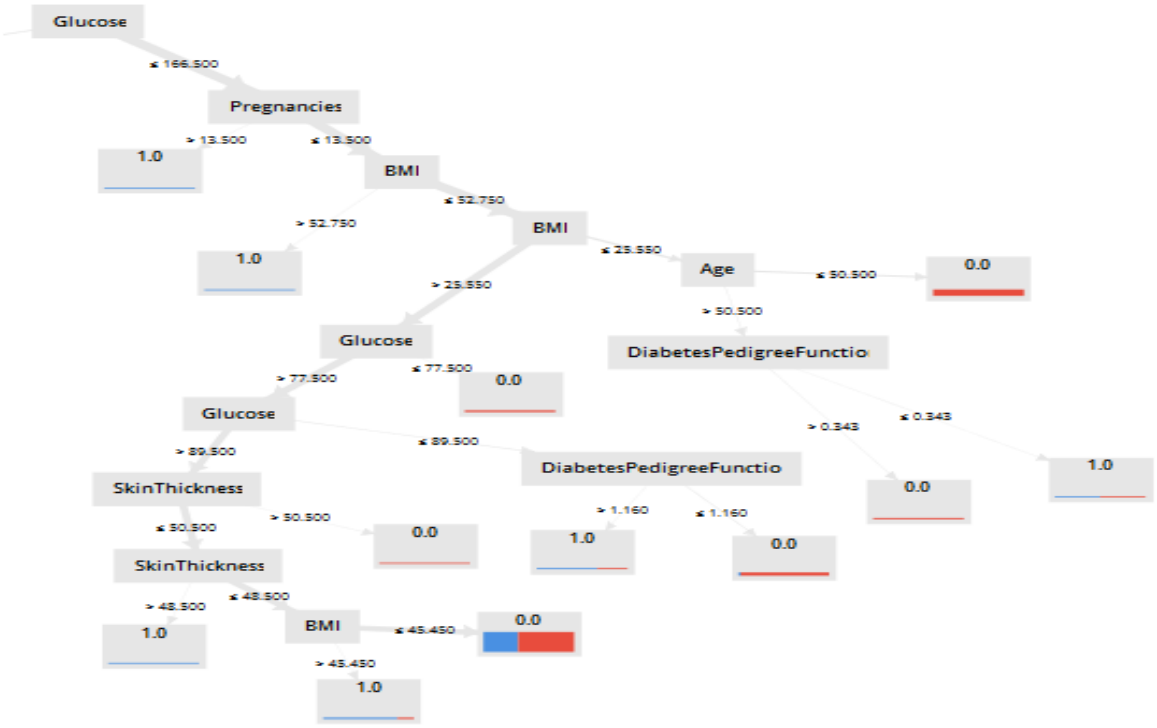
- (139+24/139+56+ 11+24)=(TP+TN/TP+FP+FN+TN)=(**accuracy**)الدقة 70.87%=
- **71.28%** =(139/139+56) =(TP/TP+FP) = **Precision**
- **92.67%** = (139/139+11) =(TP/TP+FN)= **Recall**
- **80.58%** = (2TP/2TP+FP+FN) =**F measure**

شرح شجرة القرار:

Tree

```
Glucose > 166.500
|   BloodPressure > 93
|   |   Pregnancies > 7.500: 1.0 {1.0=2, 0.0=0}
|   |   Pregnancies ≤ 7.500: 0.0 {1.0=0, 0.0=3}
|   BloodPressure ≤ 93
|   |   Pregnancies > 9.500: 1.0 {1.0=1, 0.0=1}
|   |   Pregnancies ≤ 9.500
|   |   |   DiabetesPedigreeFunction > 2.111: 1.0 {1.0=1, 0.0=1}
|   |   |   DiabetesPedigreeFunction ≤ 2.111
|   |   |   |   SkinThickness > 16: 1.0 {1.0=32, 0.0=0}
|   |   |   |   SkinThickness ≤ 16
|   |   |   |   |   DiabetesPedigreeFunction > 0.290
|   |   |   |   |   |   DiabetesPedigreeFunction > 0.329: 1.0 {1.0=8, 0.0=1}
|   |   |   |   |   |   DiabetesPedigreeFunction ≤ 0.329: 0.0 {1.0=0, 0.0=2}
|   |   |   |   |   DiabetesPedigreeFunction ≤ 0.290: 1.0 {1.0=6, 0.0=0}
Glucose ≤ 166.500
|   Pregnancies > 13.500: 1.0 {1.0=3, 0.0=0}
|   Pregnancies ≤ 13.500
|   |   BMI > 52.750: 1.0 {1.0=2, 0.0=0}
|   |   BMI ≤ 52.750
|   |   |   BMI > 25.550
|   |   |   |   Glucose > 77.500
|   |   |   |   |   Glucose > 89.500
|   |   |   |   |   |   SkinThickness > 50.500: 0.0 {1.0=0, 0.0=3}
|   |   |   |   |   |   SkinThickness ≤ 50.500
|   |   |   |   |   |   |   SkinThickness > 48.500: 1.0 {1.0=2, 0.0=0}
|   |   |   |   |   |   |   SkinThickness ≤ 48.500
|   |   |   |   |   |   |   |   BMI > 45.450: 1.0 {1.0=9, 0.0=2}
|   |   |   |   |   |   |   |   BMI ≤ 45.450: 0.0 {1.0=117, 0.0=188}
|   |   |   |   |   Glucose ≤ 89.500
|   |   |   |   |   |   DiabetesPedigreeFunction > 1.160: 1.0 {1.0=2, 0.0=1}
|   |   |   |   |   |   DiabetesPedigreeFunction ≤ 1.160: 0.0 {1.0=1, 0.0=36}
|   |   |   |   Glucose ≤ 77.500: 0.0 {1.0=0, 0.0=14}
|   |   |   BMI ≤ 25.550
|   |   |   |   Age > 50.500
|   |   |   |   |   DiabetesPedigreeFunction > 0.343: 0.0 {1.0=0, 0.0=8}
|   |   |   |   |   DiabetesPedigreeFunction ≤ 0.343: 1.0 {1.0=2, 0.0=2}
|   |   |   |   Age ≤ 50.500: 0.0 {1.0=0, 0.0=88}
```





8.4 مقارنة أداء الخوارزميات

سيقوم هذا القسم بمقارنة أداء خوارزميات التنقيب عن البيانات للانحدار الخطي وشجرة القرار بإستخدام Rapid Miner.

المعلومات التالية مفيدة للمقارنة:

1. Confusion matrix: هي جدول يستخدم غالبا لوصف آداء نموذج التصنيف.
2. الدقة (accuracy): الدقة هي مقياس للآداء و هي ببساطة نسبة الملاحظة المتوقعة بشكل صحيح .
3. Precision: هي نسبة الملاحظات الإيجابية المتوقعة بشكل صحيح إلى إجمالي الملاحظات الإيجابية المتوقعة.
4. Recall: هو نسبة الملاحظات الإيجابية المتوقعة بشكل صحيح إلى جميع الملاحظات.
5. F measure: هي المتوسط المرجح للـــ precision and recall.

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

النتائج الخاتمة و التوصيات

1.5 النتائج:

نتائج العينة				
	accuracy	**Precision**	**Recall**	**F measure**
الانحدار الخطي	%76.09	%79.14	%86.00	82.43%
شجرة القرار	70.87%	71.28%	92.67%	80.58%

1. قرائة الجدول أعلاه تبين أن ال accuracy الخاص بالإنحدار الخطي أعلى منه في شجرة القرار كذلك الـــ precision و F measure أما الـــ Recall الخاص بشجرة القرار هو الأعلى.
2. منه نستنتج أن خوارزمية الإنحدار الخطي أفضل من خوارزمية شجرة القرار في التنبوء بنوع مرض السكري.

2.5 الخاتمة:

في هذا البحث تمت دراسة إمكانية استخدام التنقيب عن البيانات للتنبؤ بمرض السكرى من النوع الاول والتاني. في التنقيب عن البيانات هناك نماذج مستخدمة في عملية التنبؤ بشكل عام إخترنا منها شجرة القرار والانحدار الخطى و قمنا بعمل مقارنة بينهم في precision،accuracy ، Recall and F measure باستخدام Rapid Miner. و إستخدمنا في هذا البحث بيانات (Pima Indians diabetics) التي تحتوي على 769 سجل و 9 خصائص .

عند تنفيذ خوارزيمة الإنحدار الخطي داخل ال Rapidminer تحصلنا على (accuracy = 76.09 %) ، (precision = 79.14%) ، (Recall = 86.00%) و (F measure = 82.43%) و عند تنفيذ شجرة القرار تحصلنا على (accuracy = 70.87 %) ، (precision = 71.28%) ، (Recall = 92.67%) و (F measure = 80.58%) مع مقارنة النتائج التي تحصلنا عليها نجد أن الإنحدار الخطي أفضل من شجرة القرار في التنبوء بنوع مرض السكري .

3.5 التوصيات:

المعوق الأساسي لهذا البحث كان جمع البيانات ، كان من الصعب الحصول على بيانات من المؤسسات الصحية في السودان لعدم توفر سجلات تحفظ بيانات المرضى (عدد مرات الحمل، الضغط ، نسبة الانسولين في الدم...الخ) لهذا اوصى الباحثين بما يلى:

1. في هذا البحث البيانات المستخدمة هي بيانات مجمَعة من الإنترنت، أوصي الباحثين بجمع بيانات حقيقية من المؤسسات الصحية بالسودان
2. السجلات التي تم جمعها كانت 769 سجل، أوصي الباحثين بجمع عدد أكبر من السجلات للحصول على نتائج أفضل
3. استخدام نماذج أخرى غير الانحدار الخطي و شجرة القرار.
4. استخدام ال (deep learning) للتنبؤ بمرض السكري .

4.5 المصادر والمراجع:

أولاً: القرآن الكريم

ثانياً: المراجع العربية:

1. جلال إبراهيم العبد، استخدام الأساليب الكمية في اتخاذ القرارات الإدارية، دار الجامعة الجديدة للنشر، 2007.
2. حمدي طه، مقدمة في بحوث العمليات، دار المريخ للنشر، د.ت.
3. نبيل محمد مرسي، الأساليب الكمية في الإدارة، المكتب الجامعي الحديث، 2006.

ثالثاً: المراجع الأجنبية:

1. (Agrawal and Srikant ،1995 استشهد بها ؛ Cooper and Schindler ،2003)
2. (Cooper and Schindler ،2003)
3. (Holme and Solvang ،1991؛ Malhotra ،2006 ؛ مالهوترا وبيترسون ، 2006)
4. (Javaheri ،2007).
5. (Javaheri, 2007)
6. (Kantardzic ،2003)
7. (Malhotra ،2006 استشهد به ؛ Javaheri ،2007)
8. (Malhotra ،2006)
9. (Malhotra and Birks ،2003 استشهد به ؛ Javaheri ،2007)
10. (Malhotra and Peterson ،2006)
11. (Pyle (199)
12. (Saunders et al. ،2000)
13. (Saunders et al. ،2000؛ Malhotra ،2006)

14. Badge, R, and p. patil , Diagnosis of Diabetes using OLAP and data mining integration . international journal of computer science &communication networks, 2012.
15. Concaro, s., l. sacchi , and R. Bellazzi. Temporal data mining for the assessment of the costs related to diabetes mellitus pharmacological treatment. In AMIA 2009.
16. Data preparing (Han and Kamber, 2006)
17. Kantardzic, M., "Data Mining: Concepts, Models, Methods, and Algorithms" John Wiley and Sons, Inc., edited by ff, 2003.
18. Mukwevho ,p. , evaluation of health and diabetes knowledge of cu 4 health participants by food frequency questionnaire and Michigan diabetes knowledge test , in food , nutrition and culinary sciences. 2010.
19. Gangil، Analysis of diabetes mellitus for early prediction using optimal features selection، Sneha and Gangil ، journal of Big Data، 2019.
20. http://armwikipedia.org