المجلة العراقية للعلوم الإحصائية

www.stats.mosuljournals.com

استخدام الجار الاقرب للمقارنة بين تصنيف العمر الحقيقي والعمر من خلال العظم لمرضى الثلاسميا

عمر فوزي صالح الراوي

قسم الادارة القانونية ، الجامعة التقنية الشمالية ، المعهد التقني نينوى ، العراق

معلومات النشر

تاريخ المقالة:
تم استلامه في 6 ايلول 2020
تم القبول في 14 تشرين الاول 2020
متاح على الإنترنت في 1 كانون الاول 2020

الكلمات الدالة:
الجار الاقرب، التصنيف، الثلاسيميا

المراسلة:
عمر فوزي صالح الراوي
omarfs@ntu.edu.iq

الخلاصة

يعتبر مرض الثلاسيميا من الامراض المزمنة ولاسيما الاطفال منذ السنوات الاولى من العمر ويمر المريض بمراحل على مدى فترات طويلة ، تم جمع البيانات للمرضى عن طريق العمر الحقيقي والعمر من خلال العظم ، لايعطي العمر من خلال العظم في فترات متقدمة من المرض صورة مقاربة لعمر المريض ، لذا سيتم عمل مقارنة بين الحالتين .

هنالك العديد من الاساليب الاحصائية المستخدمة للوصول الى تصنيف للبيانات ، تم الاعتماد على طريقة الجار الاقرب كوسيلة للتصنيف بين المجتمعات ،ان طريقة تصنيف كل مشاهدة يعتمد على اقرب ثلاث قيم يتم على اساسها وضع المشاهدة الى المجموعة الصحيحة ، ان طبيعية بيانات الدراسة كانت متقاربة نوعا ما لذا تطلب منا استخدام اسلوب يساعدنا في الوصول الى تصنيف افضل ، يعتبر الجار الاقرب هو الاسلوب الافضل للوصول الى تصنيف امثل لمثل هكذا بيانات .

كان التصنيف من خلال العمر الحقيقي افضل من التصنيف من خلال عمر العظم باستخدام تصنيف الجار الاقرب .

DOI: https://doi.org/10.33899/iqjoss.2020.167392 , ©Authors, 2020, Northern Technical University,Technical Insitute,Nineveh ,Iraq.

1- المقدمة Introduction

يعتبر التصنيف احدى الاساليب الاحصائية المستخدمة للتمييز بين عدة مستويات ، للوصول الى ترتيب للبيانات حسب مستويات ، خصوصا في العديد من الدراسات ولاسيما في الدراسات الصحية للتعرف الى اي من المستويات تنمتي اي المشاهدة ، يتم استخدام العديد من الاساليب الاحصائية للتصنيف من خلال المتغيرات ، يمكن الوصول الى قرار بالتصنيف من خلال متغير واحد او من خلال عدد من المتغيرات ، حسب الاسلوب الاحصائي وطبيعية الدراسة للوصول الى تصنيف افضل . كان هنالك العديد من الدراسات التي استخدمة اساليب مختلفة من التصنيف نذكر منها ، Chang and other (1991) استخدامهم الاساليب الوراثية في عملية التصنيف ،كما تطرق Weinberger and other (2006) الى اسلوب قياس البعد بين المشاهدات باستخدام اسلوب الجار الاقرب ،استخدم Al- Rawi (2007) دالة التميز القانونية كاسلوب في التصنيف بين الصور الرقمية ، واستخدم Al-Rawi (2007) ايضا السيطرة النوعية والدالة التمييزية في دراسة تطبيقية ، ومن المواضيع التي تم التطرق اليها من قبل Anbeek and other (2008) اسلوب تحديد تقسيمات المجاميع للشبكات العصبية من خلال استخدام تصنيف الجار الاقرب ،تم التطرق من قبل Yong and other (2009) الى اسلوب تصنيف النصوص من خلال الجار الاقرب ،كما استخدم Cai and other (2010) تصنيف الجار الاقرب لتصنيف الاوراق البحثية

،صنف Guru and other (2010) صور الزهور من خلال ملامح تركيبها باستخدام السلوب الجار الاقرب ، لم يقتصر استخدام الجار الاقرب على الطرق الكلاسيكية فقد استخدم Tomašev and other (2011) احتمالية التصنيف للجار الاقرب من خلال تحليل بيز ، استخدم Chiang and other (2012) لتصنيف متعدد المراحل من خلال الجار الاقرب ، قارن Kim and other (2012) طرق تصنيف الصور من خلال الجار الافرب والتصنيف من خلال المتجهات ،استخدم Imandoust and other (2013) تصنيف الجار الاقرب للتنبؤ بالتغيرات الاقتصادية من خلال الاعتماد على الاحداث السابقة ، كما استخدم Khamar (2013) الجار الاقرب في تصنيف النصوص القصيرة من بالاعتماد على قياس المسافة بين المجتمعات .

ان الهدف من هذا البحث هو استخدام تصنيف الجار الاقرب للمقارنة بين العمر الحقيقي للمريض والعمر من خلال العظم ، و بسبب تقارب البيانات الماخوذة للعينة ، يتم الاعتماد على هذا الاسلوب للوصول الى تصنيف افضل اي باستخدام الجار الاقرب الذي يعتمد على تصنيف القيم المتقاربة ليتم وضعها الى مجموعة معينة .

تمت كتابة البحث على مرحلتين الاولى الجانب النظري والثانية الجانب العلمي ، تم التطرق في الجانب النظري الى كيفية حساب تقسيم العينة قيد الدراسة عينة التدريب وعينة العرض ، كما ذكر طرق حساب الدول المستخدمة في التصنيف وقياس البعد بين المشاهدات بالاضافة الى ذلك تم التطرق الى ذكر اوزان المتغيرات ، اما الجانب العملي فذكر فية مما تتكون البيانات بالاضافة الى مقارنة نتائج العمر الحقيقي والعمر من خلال العظم ، من ناحية تقسم العينات التصنيفات الصحيحة واخطاء التصنيف ، رسوم بيانية توضح توزيع البيانات في الحالتين ،تم تصنيف المشاهدات الى المجتمعات ، مقارنة الاخطاء بين الحالتين ، حساب اهمية كل متغير ، ليتم الوصول الى ان العمر الحقيقي من خلال الجار الاقرب افضل بكثير من استخدام العمر من خلال العظم .

2- تقسيم العينات :

تقسيم البيانات الى عينتين الاولى التدريب والثانية العرض ، يتم استخدام عينة التدريب لتدريب نموذج يساعد الوصول الى اقرب جار، ان نسبة عينة التدريب هي 70% من اصل البيانات ، يتم الاعتماد على عينة العرض للحصول على النموذج المستخدم في التصنيف. ان نسبة عينة العرض هو 30% ، ان عينة العرض هي مجموعة مستقلة من البيانات المسجلة والتي تستخدم في الحصول على النموذج النهائي . وان الخطاء لعينة العرض يعطي تقدير "صادق" لقابلية التنبؤ للنموذج لكون عينة العرض لاتسخدم في بناء النموذج.(Reddy,2019)

3- اختيار المتغيرات:

سيتم التركيز على قوة النموذج بالاعتماد على حقل المتغيرات والتي ستعمل على ادخال واخراج المتغيرات حسب الاهمية والاعتماد على المتغيرات صاحب الاهمية الاكبر، وان لوحة اهمية المتغيرات ستساعدنا في الوصول الى اي المتغيرات الاهم ، وان لكل متغير له نسبة اقل من الواحد الصحيح ، اذا يتم جمع القيم لكل المتغيرات يكون مجموعها مساويا الى الواحد الصحيح وان اهمية المتغيرات ليس لها علاقة بدقة النموذج . وهي مرتبطة باهمية متغير التنبؤ المستخدم بالتنبؤ ، سواء التنبؤ دقيق ام لا .

بفرض ان المتغيرات $(X_1, X_2, \dots, X_m)$ $(1 \le m \le p^0)$ في النموذج لعملية الاختيار المتقدم بنسبة خطاء او مجموع مربعات الاخطاء e. ان اهمية المتغرات X_p في النموذج تحسب من الطريقة الاتية :

- يتم حذف ملامح المتغرات X في النموذج يتم التنبؤ وتقييم نسبة الخطاء او حساب اخطاء مجموعة المربعات e_p. بالاعتماد على المتغيرات $X_1, X_2, \dots X_{p-1}, X_{p+1} \dots, X_m$.
- حساب نسبة الاخطاء $e_p + \frac{1}{m}$.وتكون اهمية المتغيرات X_p هي (Wendler,2016)

$$FI_{(p)} = e_p + \frac{1}{m} \qquad 1$$

4- حساب الدالة :

بفرض ان g يمثل عدد من المجاميع ، n_iيمثل عدد المشاهدات في المجموعة i ، وان q_i هي الاحتمالية المسبقة للمجموعة i ، وافرض ان xتشير الى المشاهدة المقدرة بمتغير التمييز ذي البعد p. ولتنظيم عمل دالة التمييز x يمثل متجة المتغير.والتي تقابل صف من مجموعة البيانات وبفرض ان $f_i(x)$ تمثل دالة الكثافة الاحتمالة للمجموعة i ، وبفرض ان $P(x/G_i)$تمثل الاحتمالة الشرطية للمشاهدة x والتي تعود الى المجموعة i ، تمثل الاحتمالية المسبقة للمجموعة i والتي تعطي المشاهدة x كما في الدالة $P(G_i/x)$ ضمن نظرية بيز كما يلي (Pochiraju,2018)

$$P(G_i/x) = \frac{q_i f_i(x)}{\sum_{j=1}^{q} q_i f_i(x)} \qquad 2$$

وبتعويض $P(x/G_i)$ بدلا عن $f_i(x)$ ستكون الدالة كما في الشكل التالي

$$P(G_i/x) = \frac{q_i P(x/G_i)}{\sum_{j=1}^{q} q_i P(x/G_i)} \qquad 3$$

لدالة التمييز الجار الاقرب للحالة k نفرض ان k_i هي عدد الجار الاقرب k للمجموعة i ، وان الاحتمالية المسبقة للنموذج سيكون (McCormick,2017)

$$P(G_i/x) = \frac{\frac{q_i k_i}{n_i}}{\sum_{j=1}^{q} \frac{q_i k_i}{n_i}} \qquad 4$$

في حال وجود روابط في الجار الاقرب ، سوف تزيد k العلاقة لتتلائم مع الرابط ،اذا كان لدينا خمس نقاط متقاربة ومتساوية البعد للقيمة المعطات x ، سيتم حساب اقرب ثلاث قيم للجار الاقرب للقيمة x والتي ستساعد للحصول للجار لاقرب قيم . حدد الجار الاقرب بالاعتماد على الاختلاف او على البعد المحسوب ، ان مقياس التباعد بين للمتغير في حال المتغير المستمر وثاني القياس تم وصفة في ((((ملف قياس الاختلاف)))) ، واذا تم اختيار نفس المقياس ،والتي ستحول الى التباين باحدى الطريقتين :(IBM,2015)

$$d(ij) = \sqrt{s(ii) + s(jj) - 2s(ij)} = \sqrt{2\{1 - s(ij)\}} \qquad 5$$

$$d(ij) = 1 - s(ij) \qquad 6$$

مع اي مقياس مستمر ، تحويل مهلنوبس يفضل استخدامة قبل حساب التباين . وان اختيار قيمة k الامثل للجار الاقرب علميا ليست دقيقة . في المجموعتين ويجب ان تكون قيمة k فردية لتفادي الترابط . وان افضل قيمة للثابت k هو $\sqrt{n_i}$ حيث ان n_i يمثل الحجم النموذجي للمجموعة .

ان سبب استخدام تصنيف ال KNN عندما تكون البيانات متقاربة لذا يتم اختيار نقاط تقارب بين البيانات ، في البداية يتم اختيار اقرب جار بعدها يتم تحديد الى اي فئة تنتمي هذا القيمة . وان انضمام القيمة الى اي مجموعة يتم توقعها من خلال القيم المجاورة لها . (International,2017)

5- الجانب العلمي :

تم جمع البيانات من خلال 150 مشاهدة مصابة بمرض فقر دم البحر الابيض المتوسط نوع بيتا Beta-Thalassaemia وكانت المتغيرات كما في الجدول رقم 1 .

جدول رقم (1) يمثل اسماء المتغيرات					
ت	المتغير	اسم المتغير	ت	المتغير	اسم المتغير
1	Y1	العمر من خلال العظم	7	X6	الخلايا الشبكية
2	X1	العمر الحقيقي(شهر)	8	X7	ارومة حمراء
3	X2	العمر عند ظهور المرض	9	X8	الهيموكلوبين الجيني
4	X3	تضخم الكبد (سنتمتر)	10	X9	عدد وحدات الدم
5	X4	هيموكلبين الدم	11	X10	العمر عند اول عملية نقل للدم (شهر)
6	X5	مكدس الدم			

تم تقسيم المجاميع الى ثلاث مجاميع بموجب العمر فكانت المجموعة الاولى من العمر 1 سنة الى عمر اقل من 5 سنوات اما المجموعة الثانية من العمر 5 الى اقل من 10 سنوات والمجموعة الثالثة من العمر 10 الى 16 سنة (*Imran,2018*)،(دبدوب ، 2006)

6- ملخص المشاهدات :

من ما ذكر في الجانب النظري حول تقسيم البيانات تم تقسيم البيانات الى عينتين التدريب والعرض ،سيتم ذكر البيانات الخاصة بالعمر الحقيقي وبعدها يذكر التفاصيل البيانات للعمر من خلال العظم بالاعتماد على الجدول رقم 2 :

جدول رقم (2) يمثل ملخص البيانات							
ملخص المشاهدات بالاعتماد على العمر الحقيقي				ملخص المشاهدات بالاعتماد على العمر من خلال العظم			
		العدد	النسبة المئوية			العدد	النسبة المئوية
عينة	التدريب	103	68.70%	عينة	التدريب	115	76.70%
	العرض	47	31.30%		العرض	35	23.30%
عدد القيم المتاحة		150	100.00%	عدد القيم المتاحة		150	100.00%
المستبعدة القيم		0		القيم المستبعدة		0	
عدد القيم الكلي		150		عدد القيم الكلي		150	

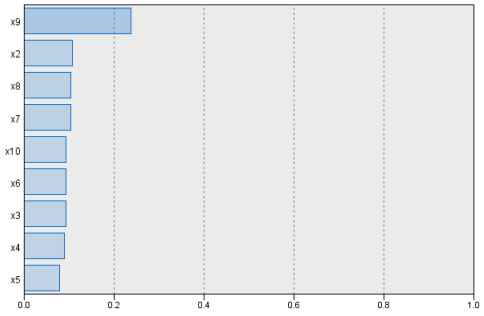
العمر الحقيقي : كان نسبة عينة التدريب هي 68.70% والبالغ عددها 103 مشاهدة ، كان نسبة عينة العرض هي 31.1 % والبالغ عددها 47 مشاهدة ، ان المجموع الكلي للمشاهدات التدريب والعرض 150 مشاهدة لتكون النسبة المئوية 100 % ولايوجد مشاهدات مستبعدة . العمر من خلال العظم : كان نسبة عينة التدريب هي 76.70% والبالغ عددها 115 مشاهدة ، كان نسبة عينة العرض هي 23.30 % والبالغ عددها 35 مشاهدة ، ان المجموع الكلي للمشاهدات التدريب والعرض 150 مشاهدة لتكون النسبة المئوية 100 % ولايوجد مشاهدات مستبعدة .

اهمية المتغيرات:

سيتم ذكر اهمية المتغيرات بالاضافة الى وزن كل متغير بالاعتماد على الرسم البياني فضلا عن الجدول المرفق مع الرسم والذي يبين اهمية كل متغير بالنسبة الى العمر الحقيقي للمريض والعمر من خلال العظم من خلال العمر الحقيقي وبالاعتماد على الجدول رقم 3 والشكل رقم 1 والذي يمثل المدرج التكراري ،تم توضيح اهمية المتغيرات حسب وزن كل متغير فكانت المتغيرات حسب الاهمية من اكثر متغير مهم الى اقل متغير مهم على النحو التالي فان المتغير الاكثر اهمية هو (x9=0.2375) والمتغير الثاني حسب الاهمية هو (x2=0.1072) المتغير الثالث حسب الاهمية هو (x7=0.1037) المتغير الرابع حسب الاهمية هو (x8=0.1037) المتغير الخامس حسب الاهمية (x3=0.0931) المتغير السادس حسب الاهمية هو (x6=0.0931) المتغير السابع حسب الاهمية هو (x10=0.0931) المتغير الثامن حسب الاهمية هو (x4=0.0896) المتغير التاسع حسب الاهمية هو (x5=0.079) . من خلال ما ذكر اعلاه فانه سيتم اختيار اول ثلاث متغيرات وتهمل البقية في الرسم وسيتم الاعتماد على المتغيرات التالية حسب التسلسل(x7,x2,x9) .

جدول رقم (4) يمثل اهمية المتغيرات
للعمر الحقيقي

عينة التدريب	
المتغير	اهمية المتغير
x9	0.2375
x2	0.1072
x7	0.1037
x8	0.1037
x3	0.0931
x6	0.0931
x10	0.0931
x4	0.0896
x5	0.079

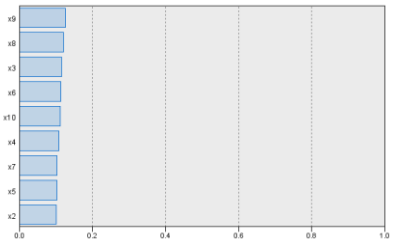


شكل رقم (1) يمثل مدرج تكراري للمتغيرات حسب الاهمية للعمر الحقيقي

من خلال العمر من خلال العظم وبالاعتماد على الجدول رقم 4 والشكل رقم 2 والذي يمثل المدرج التكراري ،تم توضيح اهمية المتغيرات حسب وزن كل متغير فكانت المتغيرات حسب الاهمية من اكثر متغير مهم الى اقل متغير مهم على النحو التالي ،ان المتغير الاكثر اهمية هو (x9=0.126) والمتغير الثاني حسب الاهمية هو (x8=0.1203) المتغير الثالث حسب الاهمية هو (x3=0.1147) المتغير الرابع حسب الاهمية هو (x6=0.1128) المتغير الخامس حسب الاهمية

جدول رقم (4) يمثل اهمية المتغيرات
للعمر من خلال العظم

عينة العرض	
المتغير	اهمية المتغير
x9	0.126
x8	0.1203
x3	0.1147
x6	0.1128
x10	0.1109
x4	0.1071
x5	0.1034
x7	0.1034
x2	0.1015

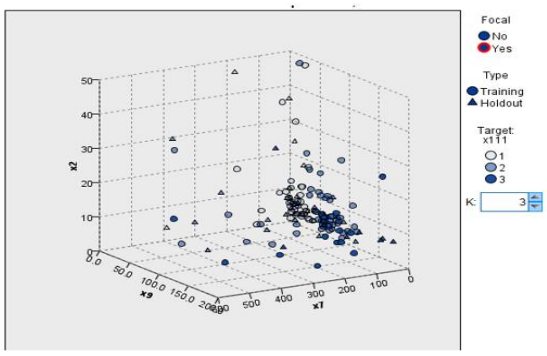


شكل رقم (2) يمثل مدرج تكراري للمتغيرات حسب الاهمية للعمر من خلال العظم

(x10=0.1109) المتغير السادس حسب الاهمية هو (x4=0.1071) المتغير السابع حسب الاهمية هو (x5=0.1034) المتغير الثامن حسب الاهمية هو (x7=0.1034) المتغير التاسع حسب الاهمية هو (x2=0.1015) . من خلال ما ذكر اعلاه فانه سيتم اختيار اول ثلاث متغيرات وتهمل البقية في الرسم وسيتم الاعتماد على المتغيرات التالية حسب التسلسل(x3,x8,x9) ، الرسم ثلاثي الابعاد.

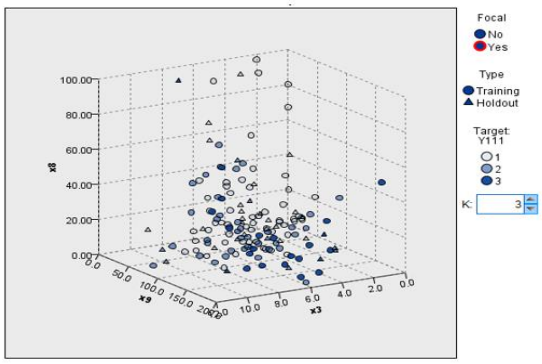
7- الرسم البياني للمجاميع :

من الشكل رقم 3 والذي يمثل الرسم البياني ثلاثي الابعاد ،تم استخدام الرسم البياني ثلاثي الابعاد بالاعتماد على ثلاث متغيرات (x7,x2,x9) ان هذه المتغيرات الاكثر اهمية وتم اهمال بقية المتغيرات اي تم تصنيف البيانات بموجب الرسم ثلاثي الابعاد ، ولتوضيح نفس الرسم فقد تم تلوين المجاميع الى ثلاث الوان مختلف مؤشرة على جانب الرسم (1،2،3)، من خلال ما ذكر اعلاه فان البيانات كانت على نوعين التدريب والعرض اخذت بيانات التدريب شكل المثلث بالنسبة الى المجموعات الثلاث اما بيانات العرض فاخذ شكل الدائرة بالنسبة الى بيانات العرض .



شكل رقم (3) يمثل رسم بياني لتوزيع المشاهدات حسب المجموعات لعينتين التدريب والعرض للعمر الحقيقي

من الشكل رقم 4 والذي يمثل الرسم البياني ،تم استخدام الرسم البياني ثلاثي الابعاد بالاعتماد على ثلاث متغيرات (x3,x8,x9) ان هذه المتغيرات الاكثر اهمية وتم اهمال بقية المتغيرات اي تم تصنيف البيانات بموجب الرسم ثلاثي الابعاد ، ولتوضيح نفس الرسم فقد تم تلوين المجاميع الى ثلاث الوان مختلف مؤشرة على جانب الرسم (3،2،1) ، من خلال ما ذكر اعلاه فان البيانات كانت على نوعين التدريب والعرض اخذت بيانات التدريب شكل المثلث بالنسبة الى المجموعات الثلاث اما بيانات العرض فاخذ شكل الدائرة بالنسبة الى بيانات العرض .

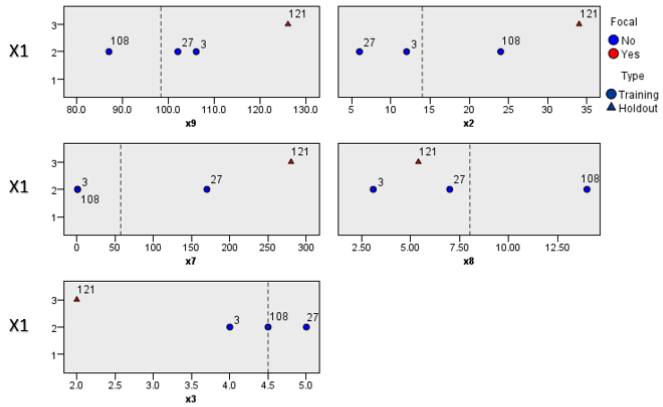


شكل رقم (4) يمثل رسم بياني لتوزيع المشاهدات حسب المجموعات لعينتين التدريب والعرض للعمر من خلال العظم

8- اختيار الجار الاقرب:

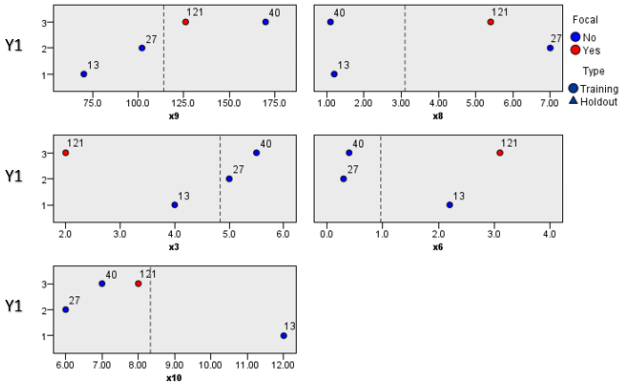
تم اعتماد قيمة المتغير k=3 اي سيكون اتخاذ القرار بالنسبة الى الجار الاقرب من خلال اقرب ثلاث مشاهدات ، تم اختيار نقطة بشكل عشوائي للتوضيح ، فاعطت اقرب ثلاث نقاط وهو الاسلوب المستخدم في اختيار المجموعة بالنسبة الى الجار الاقرب لثلاث قيم ، باستخدام اسلوبين الاول من خلال رسم المتغيرات وكما هو موضح في الرسمين التوضيحيين (1) و(2) ، والطريقة الثانية من خلال جدول لقياس البعد بين المشاهدات كما في الجدولين (5) و(6) .

العمر الحقيقي :من الشكل رقم 5 والذي يمثل الرسم التوضيحي ،تم توضح الجار الاقرب بالنسبة الى قيمة المشاهدة المختارة وهي 121 وسيتم اختيار الجار الاقرب من خلال القيم القريبة من خلال عدة متغيرات بالنسبة الى متغير العمر الحقيقي(X1) بالاضافة الى المتغيرات الاهم فالاقل اهمية ، فكانت المشاهدات (3،108،27) هي الجار الاقرب للمشاهدة 121 ويمكن مشاهدتها حسب المتغيرات (x3,x8,x7,x2,x9) والتي بدورها ستنظم الى المجموعة الثانية من خلال قرب القيمة الى قيم المجموعة الثانية .



شكل رقم (5) يمثل رسم توضيحي للجار الاقرب بالنسبة الى اقرب مشاهدات بالاعتماد على العمر الحقيق

العمر من خلال العظم : من الشكل رقم 6 والذي يمثل الرسم التوضيحي ،تم توضح الجار الاقرب بالنسبة الى قيمة المشاهدة المختارة وهي 121 وسيتم اختيار الجار الاقرب من خلال القيم القريبة من خلال عدة متغيرات بالنسبة الى متغير العمر من خلال العظم (Y1) بالاضافة الى المتغيرات الاهم فالاقل اهمية ، فكانت المشاهدات (13،40،27) هي الجار الاقرب للمشاهدة 121 ويمكن مشاهدتها حسب المتغيرات (x10,x6,x3,x8,x9) والتي بدورها ستنظم الى المجموعة الثانية من خلال قرب القيمة الى قيم المجموعة الثالثة .وبالاعتماد على هذا الاسلوب تم تصنيف المشاهدات الى المجاميع.



شكل رقم (6) يمثل رسم توضيحي للجار الاقرب بالنسبة الى اقرب مشاهدات بالاعتماد على العمر من خلال العظم

جدول رقم (5) يمثل الجار الاقرب ذي البعد (k=3) وقياس البعد لعينة التدريب						
عرض القيم الاقرب لمركز المشاهدة 121						
قيمة المركز	مشاهدات الجار الاقرب			الاقرب مسافة		
	1	2	3	1	2	3
121	27	108	3	0.493	0.493	0.502

جدول رقم (6) الجار الاقرب ذي البعد (k=3) وقياس البعد لعينة العرض						
عرض القيم الاقرب لمركز المشاهدة 121						
قيمة المركز	مشاهدات الجار الاقرب			الاقرب مسافة		
	1	2	3	1	2	3
121	27	40	13	0.49	0.491	0.492

العمر الحقيقي : ومن خلال الجدول رقم (5) الخاص بقيم الجار الاقرب والذي بين الجار الاقرب للمشاهدة 121 مع المشاهدات الاقرب ذات التسلسل (3،108،27) ذات الابعاد والتي تمثل القيم الاقرب للمشاهدة المذكورة لذا يمكن القول ان قيم الابعاد(0.493،493،0.502) يمكن الاعتماد عليها لتصنف القيمة 121 الى المجموعة الثانية بالاعتماد على الجار الاقرب حسب الجدول رقم (5) .

العمر من خلال العظم : ومن خلال الجدول رقم (6) الخاص بقيم الجار الاقرب والذي بين الجار الاقرب للمشاهدة 121 مع المشاهدات الاقرب ذات التسلسل (13،40،27) ذات الابعاد والتي تمثل القيم الاقرب للمشاهدة المذكورة لذا يمكن القول ان قيم الابعاد(0.490،0.491،0.492) يمكن الاعتماد عليها لتصنف القيمة 121 الى المجموعة الثالثة بالاعتماد على الجار الاقرب حسب الجدول رقم (6) .

جدول رقم (7) يمثل التصنيف الصحيح واخطاء التصنيف

الجزء	المشاهدة	جدول التصنيف للعمر الحقيقي				جدول التصنيف لعمر العظم			
		التنبؤ				التنبؤ			
		1	2	3	التصنيف نسبة الصحيح	1	2	3	التصنيف نسبة الصحيح
التدريب	1	34	4	0	89.50%	54	4	1	91.50%
	2	3	49	1	92.50%	20	12	5	32.40%
	3	0	3	9	75.00%	9	6	4	21.10%
	النسبة من المجموع الكلي	35.90%	54.40%	9.70%	89.30%	72.20%	19.10%	8.70%	60.90%
العرض	1	15	3	0	83.30%	18	5	0	78.30%
	2	5	17	0	77.30%	5	1	2	12.50%
	3	0	3	4	57.10%	3	0	1	25.00%
	النسبة من المجموع الكلي	42.60%	48.90%	8.50%	76.60%	74.30%	17.10%	8.60%	57.10%

9- جدول التصنيف :

حسب ما ذكر اعلاه حول جدول التقسيم للبيانات الى عينتين التدريب والعرض ، فقد تم عمل جدول لعرض التصنيف الصحيح واخطاء التصنيف للعينتين التدريب والعرض للعمر الحقيقي والعمر من خلال العظم وكانت على النحو التالي :

عينة التدريب : ان عدد المشاهدات التي تنتمي الى المجموعة الاولى وهي تنتمي الى المجموعة الاولى هي 34 مشاهدة ، اما عدد المشاهدات التي تنتمي الى المجموعة الاولى وصنفت بشكل خاطى الى المجموعة الثانية فكانت 4 مشاهدات ولاتوجد مشاهدة صنفت الى المجموعة الثالثة بشكل خاطى ، ان نسبة التصنيف في المجموعة الاولى بشكل صحيح هو 89.5% اما فيما يخص العمر من خلال العظم فكانت عدد المشاهدات التي تنتمي الى المجموعة الاولى وهي تنتمي الى المجموعة الاولى هي 54 مشاهدة ، اما عدد المشاهدات التي تنتمي الى المجموعة الاولى وصنفت بشكل خاطى الى المجموعة الثانية فكانت 4 مشاهدات ومشاهدة واحدة تنتمي المجموعة الثالثة وصنفت الى المجموعة الثالثة وهي تنتمي الى المجموعة الاولى ، ان نسبة التصنيف في المجموعة الاولى بشكل صحيح هو 91.5% ،. اما بالنسبة الى العمر الحقيقي المجموعة الثانية تم تصنيف ال3 مشاهدات الى المجموع الاولى وهي تنتمي الى المجموعة الثانية ، وتصنيف 49 مشاهدة بشكل صحيح الى المجموعة الثانية وهي منتمية الى المجموعة الثانية ، تصنيف مشاهدة الى المجموعة الثالثة وهي تنتمي الى المجموعة الثانية ، ان نسبة التصنيف الصحيح للمجموعة الثانية هي 92.5% ، اما بالنسبة الى العمر من خلال العظم ففي المجموعة الثانية تم تصنيف 20 مشاهدات الى المجموع الاولى وهي تنتمي الى المجموعة الثانية ، وتصنيف 12 مشاهدة بشكل صحيح الى المجموعة الثانية وهي منتمية الى المجموعة الثانية ، تصنيف 5 مشاهدات الى المجموعة الثالثة وهي تنتمي الى المجموعة الثانية ، ان نسبة التصنيف الصحيح للمجموعة الثانية هي 32.4% ، اما بالنسبة الى العمر الحقيقي المجموعة الثالثة لم تصف اي مشاهدة الى المجموعة الاولى ، تم تصنيف 3 مشاهدات الى المجموعة الثانية وهي تنتمي الى المجموعة الثالثة ، وتصنيف 9 مشاهدات الى المجموعة الثالثة وهي تنتمي الى المجموعة الثالثة ، ان نسبة التصنيف الصحيح للمجموعة الثالثة هو 75% وهي اقل نسبة تصنيف صحيح في عينة التدريب للعمر الحقيقي، اما بالنسبة الى العمر من خلال العظم ففي المجموعة الثالثة تم تصنيف 9 مشاهدات الى المجموعة الاولى وهي تنتمي الى المجموعة الثالثة ، تم تصنيف 6 مشاهدات الى المجموعة الثانية وهي تنتمي الى المجموعة الثالثة ، وتصنيف 4 مشاهدات الى

المجموعة الثالثة وهي تنتمي الى المجموعة الثالثة ، ان نسبة التصنيف الصحيح للمجموعة الثالثة هو 21.1% وهي اقل نسبة تصنيف صحيح في عينة التدريب ، للعمر من خلال العظم .

عينة العرض : بالنسبة الى العمر الحقيق ان عدد المشاهدات التي تنتمي الى المجموعة الاولى وهي تنتمي الى المجموعة الاولى هي 15 مشاهدة ، اما عدد المشاهدات التي تنتمي الى المجموعة الاولى وصنفت بشكل خاطى الى المجموعة الثانية فكانت 3 مشاهدات ولاتوجد مشاهدة صنفت الى المجموعة الثالثة بشكل خاطى ، ان نسبة التصنيف في المجموعة الاولى بشكل صحيح هو 83.3% ، اما بالنسبة الى العمر من خلال العظم فكانت عدد المشاهدات التي تنتمي الى المجموعة الاولى وهي تنتمي الى المجموعة الاولى هي 18 مشاهدة ، اما عدد المشاهدات التي تنتمي الى المجموعة الاولى وصنفت بشكل خاطى الى المجموعة الثانية فكانت 5 مشاهدات ولاتوجد مشاهدة صنفت الى المجموعة الثالثة بشكل خاطى ، ان نسبة التصنيف في المجموعة الاولى بشكل صحيح هو 78.3% . اما بالنسبة الى العمر الحقيقي المجموعة الثانية تم تصنيف 5 مشاهدات الى المجموع الاولى وهي تنتمي الى المجموعة الثانية ، وتصنيف 17 مشاهدة بشكل صحيح الى المجموعة الثانية وهي منتمية الى المجموعة الثانية ، ولا يوجد تصنيف الى المجموعة الثالثة وهي تنتمي الى المجموعة الثانية ، ان نسبة التصنيف الصحيح للمجموعة الثانية هي 77.3% ،اما بالنسبة الى العمر من خلال العظم ففي المجموعة الثانية تم تصنيف 5 مشاهدات الى المجموع الاولى وهي تنتمي الى المجموعة الثانية ، وتصنيف 1 مشاهدة بشكل صحيح الى المجموعة الثانية وهي منتمية الى المجموعة الثانية ، تم تصنيف 2 مشاهدة الى المجموعة الثالثة وهي تنتمي الى المجموعة الثانية ، ان نسبة التصنيف الصحيح للمجموعة الثانية هي 12.5%، اما بالنسبة الى العمر الحقيقي المجموعة الثالثة لم تصف اي مشاهدة الى المجموعة الاولى ، تم تصنيف 3 مشاهدات الى المجموعة الثانية وهي تنتمي الى المجموعة الثالثة ، وتصنيف 4 مشاهدات الى المجموعة الثالثة وهي تنتمي الى المجموعة الثالثة ، ان نسبة التصنيف الصحيح للمجموعة الثالثة هو 57.1% وهي اقل نسبة تصنيف صحيح في عينة التدريب بالنسبة الى العمر الحقيقي ،اما بالنسبة الى العمرمن خلال العظم ففي المجموعة الثالثة تم تصنيف 3 مشاهدات الى المجموعة الاولى وهي تنتمي الى المجموعة الثالثة ، لم يتم تصنيف اي مشاهدة الى المجموعة الثانية بشكل خاطى ، وتصنيف 1 مشاهدة الى المجموعة الثالثة وهي تنتمي الى المجموعة الثالثة ، ان نسبة التصنيف الصحيح للمجموعة الثالثة هو 25.0% .

من حالتين التدريب والعرض تبين ان نسبة التصنيف الصحيح في عينة التدريب افضل بكثير من مجموعة العرض وكما هو مبين في جدول التصنيف للتدريب والعرض من خلال ما ذكر في الجدولين ان تصنيف البيانات من خلال العمر الحقيقي افضل من تصنيف البيانات من خلال عمر العظم للمريض وحسب الفروقات الواضحة في الجدول رقم (7) .

10- ملخص الاخطاء:

جدول رقم (8) يمثل ملخص الاخطاء

	للعمر الحقيقي	للعمر من خلال العظم
الجزء	**النسبة المئوية لاخطاء التصنيف**	**النسبة المئوية لاخطاء التصنيف**
التدريب	10.70%	39.10%
العرض	23.40%	42.90%

الجدول رقم (8) والذي يمثل نسبة الخطاء للعمر الحقيقي للمريض ، ان نسبة الخطاء في عينة التدريب هي 10% اما بالنسبة الى عينة العرض فان النسبة كبيرة جدا تصل الى 23% وهي نسبة كبيرة مقارنة بعينة التدريب .

من الجدول رقم (8) والذي يمثل نسبة الخطاء للعمر من العظم للمريض ، ان نسبة الخطاء في عينة التدريب هي 39.10% اما بالنسبة الى عينة العرض فان النسبة كبيرة جدا تصل الى 42.90% وهي نسبة كبيرة مقارنة بالتدريب .

11- النتائج التوصيات :

في البداية سيتم مقارنة النتائج بين عينتين العمر الحقيقي والعمر من خلال العظم وعلى مرحلتين عينة التدريب وعينة العرض النحو التالي:

عينة التدريب : ان ما ذكر بالمقارنة بين العمر الحقيقي والعمر من خلال العظم يمكن القول بان ان الدقة في المجموعة الاولى كانت التصنيف في العمر من خلال العظم 91.5 % افضل من العمر الحقيقي 89.5% لكون المرض في بدايتة ، كما ان المرض لم يوثر العمر من خلال العظم.

اما في المرحلة الثانية التي بدا المرض بالتاثير على المريض نلاحظ ان التصنيف من خلال العمر الحقيقي 92.5%هو افضل بكثير من العمر من خلال العظم 32.1% وذلك بسبب تاثير المرض على المصابين .

كذلك يمكن ملاحظة الفرق الكبير في التصنيف من خلال المجموعة الثالثة ايضا حيث ان التصنيف الصحيح للعمر الحقيقي هو 75% مقارنة مع العمر من خلال العظم 21.1% .

عينة العرض : ان ما ذكر بالمقارنة بين العمر الحقيقي والعمر من خلال العظم يمكن القول بان ان الدقة في المجموعة الاولى كانت التصنيف في العمر الحقيقي 83.3 % افضل من العمر من خلال العظم 87.3% .

اما في المرحلة الثانية التي بدا المرض بالتاثير على المريض نلاحظ ان التصنيف من خلال العمر الحقيقي 77.3%هو افضل بكثير من العمر من خلال العظم 12.5% وذلك بسبب تاثير المرض على المصابين .

كذلك يمكن ملاحظة الفرق الكبير في التصنيف من خلال المجموعة الثالثة ايضا حيث ان التصنيف الصحيح للعمر الحقيقي هو 57.1% مقارنة مع العمر من خلال العظم 25% . ان دقة التصنيف في عينتين التدريب والعرض للعمر الحقيق افضل بكثير من العمر من خلال العظم .

اما بالنسبة الى اجمالي الاخطاء في عينتي التدريب والعرض فكانت عينة التدريب افضل بشكل واضح مقارنتة بعينة العرض وان نسبة الخطاء اقل بشكل اوضح في عينتي التدريب والعرض للعمر الحقيقي مقارنة بالعمر من خلال العظم .

Refference

1- Dabdoub, Marwan Abdel-Aziz and Al-Nuaimi, Aswan Muhammad Tayeb (2006), “Suggested Methods for Character Decline”, Iraqi Journal of Statistical Sciences, University of Mosul, Iraq, 85-106Al-Rawi.O.F ,(2007), The Discrimination between Digital Photos by Using Canonical Discriminate Function, TANMIAT AL-RAFIDAIN, 29(88),221-233.

2- Al-Rawi.O.F, (2007). The use of quality control and the discriminatory function in applied studies .Journal of education and Science ,19(17),203-220.

3- Anbeek, Petronella, Koen L. Vincken, and Max A. Viergever. (2008). "Automated MS-lesion segmentation by k-nearest neighbor classification." MIDAS Journal

4- Cai, Yun-lei, Duo Ji, and Dongfeng Cai. (2010) "A KNN Research Paper Classification Method Based on Shared Nearest Neighbor." NTCIR..

5- Chiang, Tsung-Hsien, Hung-Yi Lo, and Shou-De Lin. (2012) "A ranking-based KNN approach for multi-label classification." Asian Conference on Machine Learning...

6- Chang, Eric I., and Richard P. Lippmann. (1991) "Using genetic algorithms to improve pattern classification performance." Advances in neural information processing systems..

7- IBM Corporation. (2011). IBM SPSS Modeler 14 . 2 User ’s Guide.

8- Guru, D. S., Y. H. Sharath, and S. Manjunath. (2010)"Texture features and KNN in classification of flower images." IJCA, Special Issue on RTIPPR (1) : 21-29.

9- IBM. (2015). IBM SPSS Modeler 17 Algorithms Guide. IBM SPSS Modeler 17 Algorithms Guide, 73–86.

10- International Business Machines Corporation. (2017). IBM SPSS Modeler 18.1.1 Algorithms Guide. 806.

11- Imandoust, Sadegh Bafandeh, and Mohammad Bolandraftar. (2013)"Application of k-nearest neighbor (knn) approach for predicting economic events: Theoretical background." International Journal of Engineering Research and Applications 3.5: 605-610.

12- Khamar, Khushbu. (2013)"Short text classification using kNN based on distance function." International Journal of advanced research in computer and communication engineering 2.4: 1916-1919

13- Kim¹, J. I. N. H. O., Kim, B. S., & Savarese, S. (2012). Comparing image classification methods: K-nearest-neighbor and support-vector-machines. In Proceedings of the 6th WSEAS international conference on Computer Engineering and Applications, and Proceedings of the 2012 American conference on Applied Mathematics (Vol. 1001, pp. 48109-2122).

14- Imran, A., Wani, G., & Singh, K. (2018). Assessment of thyroid profile in children with thalassemia and its correlation with serum ferritin level. Assessment, 4(10).

15- McCormick, Keith, and Jesus Salcedo. (2017) IBM SPSS Modeler essentials: Effective techniques for building powerful data mining and predictive analytics solutions. Packt Publishing Ltd,.

16- Pochiraju, Bhimasankaram, and Sridhar Seshadri, eds. (2018) Essentials of Business Analytics: An Introduction to the Methodology and Its Applications. Vol. 264. Springer,.

17- Reddy, M. Venkataswamy. (2019)Statistical methods in psychiatry research and SPSS. CRC Press.

18- Tomašev, Nenad, et al. (2011)"A probabilistic approach to nearest-neighbor classification: Naive hubness bayesian knn." Proc. 20th ACM Int. Conf. on Information and Knowledge Management (CIKM).

19- Weinberger, Kilian Q., John Blitzer, and Lawrence K. Saul. (2006)"Distance metric learning for large margin nearest neighbor classification." Advances in neural information processing systems.

20- Wendler, Tilo, and Sören Gröttrup. (2016)Data mining with SPSS modeler: theory, exercises and solutions. Springer.

21- Yong, Zhou, Li Youwen, and Xia Shixiong. (2009)"An improved KNN text classification algorithm based on clustering." Journal of computers 4.3: 230-237.

Use the k nearest neighbor(KNN) to compare the classification of real age and age through the bone for thalassic patients

Omar Fawzi Saleh Al-Rawi
Department of Legal Administration, Northern Technical University, Nineveh Technical Institute, Iraq

Abstract:

Thalassemia is considered a chronic disease, especially children from the first years of life, and the patient goes through stages over long periods, Data were collected for patients by real age and age through the bone, Therefore, a comparison will be made between the two cases.

There are many statistical methods used to arrive at a classification of data, the method of nearest neighbor has been relied upon as a method of classification between societies. The method of classifying each observation depends on the three closest values on the basis of which the observation is placed into the correct group, the naturalness of the data was rather close, so it asked us to use a method that helps us to reach a better classification. The k the nearest neighbor is the best way to reach an optimal classification for such data. Classification by real age was better than classification by bone age using classification. Classification by actual age was better than classification by bone age using k nearest neighbor classification

Keyword: Nearest neighbor, classification, thalassemia.