4- 1 تمهيد:-

في هذا الفصل سيتم التطرق علي طريقتي التحليل العنقودي موضع الدراسة ، حيث يتم تطبيق الطريقتين علي مجموعة بيانات (Data Set) ، باستخدام برنامج R ،حيث تعد لغة R من اللغات التي صعد نجمها حديثا وبشكل سريع بمجال البرمجة العلمية في قطاعي الإحصاء والمعلوماتية الحيوية (Bioinformatics) .

حيث باتت معتمدة على نطاق واسع في كثير من الجامعات ومراكز البحث العلمية، وأصبحنا نرى استخدامها والإشارة إليها في المقالات المنشورة بالمجلات العلمية المحكّمة يزداد بشكل طردي ومتسارع، هذا عدى عن حقيقة كونها لغة حرة مفتوحة المصدر يخضع توزيعها لترخيص GPL الشهير.[1]

جامعة كاليفورنيا هي جامعة بحثية عامة تقع في إرفين، كاليفورنيا، الولايات المتحدة الأمريكية، وتعتبر واحدة من 10 جامعات في نظام جامعات كاليفورنيا[2].

تم استخدام بيانات مجهزة لاغراض البحث العلمي من موقع مركز تعليم الآلة والأنظمة الذكية[3] التابع للجامعة حيث يوفر الموقع بيانات تم جمعها من المصادر المختلفة وإتاحتها للباحثين العاملين في نطاق تعليم الآلة والتحليل الاحصائي وتوليد البيانات , تم البدء في تكوين هذا الارشيف عام 1987 بواسطة "David Aha" وطلابه من الجامعة .

منذ ذلك الحين تم استخدام ارشيف البيانات هذا علي نطاق واسع حول العالم من قِبل الباحثين والطلاب باعتبارة المصدر الرئيسي لمجموعات البيانات لاغراض تعليم الآلة "Machine Learning" وقد تم تصنيفة والاستشهاد به أكثر من 1000 مرة في الاوراق العلمية لذلك صنف الموقع من ضمن أفضل 100 موقع والاكثر ذكراً حول العالم في مجال الحاسوب , تم تصميم الاطار الحالي للموقع عام 2007 بواسطة " Arthur Asuncion" و David Newman" وقام هذا المشروع بالتعاون مع "Rexa.info" في جامعة ماساتشوستس أمهرست .[4]

فيما يلي جدول رقم (1) يوضح المتغيرات التي ادخلت في البحث وتعريفها [5] :-

4-2 تعريف ووصف متغيرات الدراسة:-

فيما يلي انواع المتغيرات التي ادخلت في الدراسة ووصف كل متغير مع عمود يوضح نسبة القيم المفقودة في بيانات الدراسة لكل متغير.

[1] https://academy.hsoub.com/programming/r-language
[2] https://www.universityofcalifornia.edu/ 20/11/22016 - 22:00 pm
[3] https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008 15/11/2016 - 08:30am
[4] https://archive.ics.uci.edu/ml/about.html 20/11/2016 - 22:00 PM
[5] Bio-Med Research International Volume 2014 (2014), Article ID 781670, 11 pages
https://www.hindawi.com/journals/bmri/2014/781670/tab1/

جدول (4–1) متغيرات الدراسة

اسم المتغير	وصف المتغير	القيم المفقودة %
السلالة	سلالة المريض	2
النوع	نوع المريض بالاضافة الي تعريف عدم تحديد النوع كقيمة غير معلومة	0
العمر	عمر المريض مقسم في 10 فئات بطول 10 سنوات في كل فئة من 0 – 100	0
الفترة الزمنية	فترة اقامة المريض وحتي خروجه في المدي بين (1 – 14) يوم	0
الإجراءات المعملية	عدد الاختبارات المعملية التي اجريت للمريض	0
الاجراءات	عدد الاجراءات (غير المعملية) التي قام بها المريض اثناء اقامته في المستشفي	0
الادوية	عدد الادوية التي قام بأخذها	0
العيادات الخارجية	عدد الزيارات للعيادات الخارجية خلال عام فعلياً	0
الزيارات الطارئة	عدد الزيارات الطارئة التي قام بها المريض خلال العام	0
الزيارات الداخلية	عدد الزيارات للمستشفي	0
التشخيصات	عدد التشخيصات الطبية التي ادخلت الي النظام	0

المصدر : اعداد الباحث باستخدام برنامج MS-Word

4–3 وصف المتغيرات:

اولاً : السلالة:–

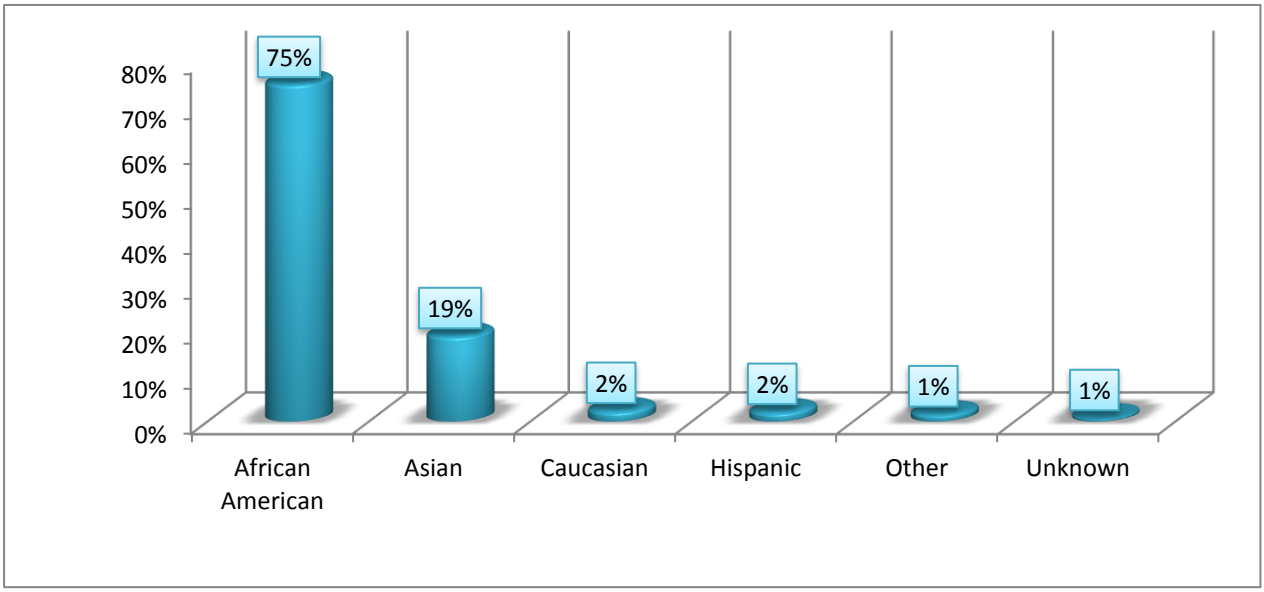
جدول (4–2) وصف متغير السلالة

السلالة	التكرار	النسبة %
امريكيون من اصول افريقية African American	76099	75
آسيويون Asian	19210	19
قوقازييون Caucasian	2273	2
من اصل اسباني Hispanic	2037	2
من اصول اخري او مختلطة Other	1506	1
غير معلومي الاصل Unknown	641	1
المجموع	101766	100.00

المصدر : اعداد الباحث باستخدام برنامج MS-Excel

الجدول (4–2) توزيع اعداد المرضي حسب سلالتهم حيث نجد ان السلالة الافريقية المهاجرة الي امريكا (African American) تمثل (75%) من مجموع السلالات الكلي ويوضح الشكل (4–1) نسب كل سلالة من سلالات المرضي في حزمة البيانات ونلاحظ ان السلالة التي تليها هي الآسيويون (Asian) بنسبة بلغت (19%).

الشكل (4–1) وصف متغير السلالة



المصدر : اعداد الباحث باستخدام برنامج MS-Excel

ثانيا : النوع:

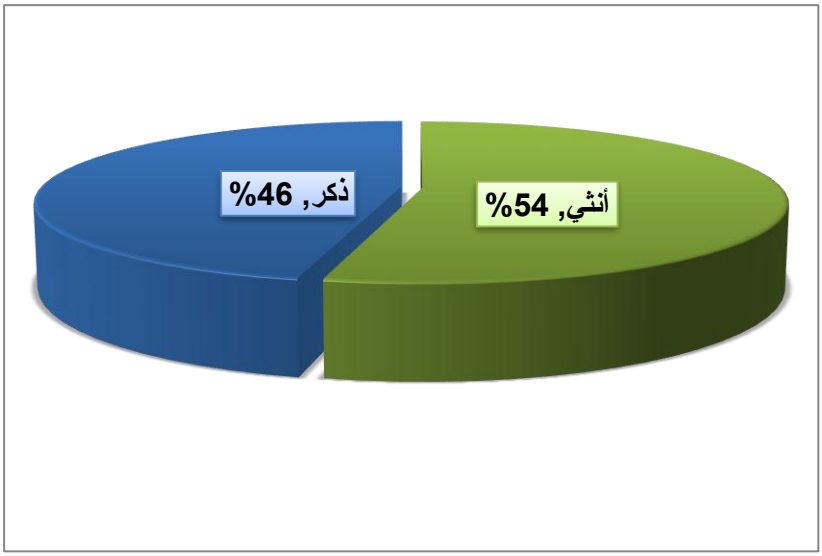
جدول (4–3) نوع المريض

النوع	التكرار	النسبة %
أنثي	54,708	53.76
ذكر	47,055	46.24
غير محدد	3	0.00
المجموع	101766	100.00

المصدر : اعداد الباحث باستخدام برنامج MS-Excel

الجدول (4–3) يوضح توزيع اعداد المرضي حسب النوع ومن الشكل (4–2) يتضح ان نسبة اصابة الاناث (54%) تفوق بقليل نسبة الذكور.

الشكل (4–2) نوع المريض



المصدر : اعداد الباحث باستخدام برنامج MS-Excel

ثالثا : العمر:–

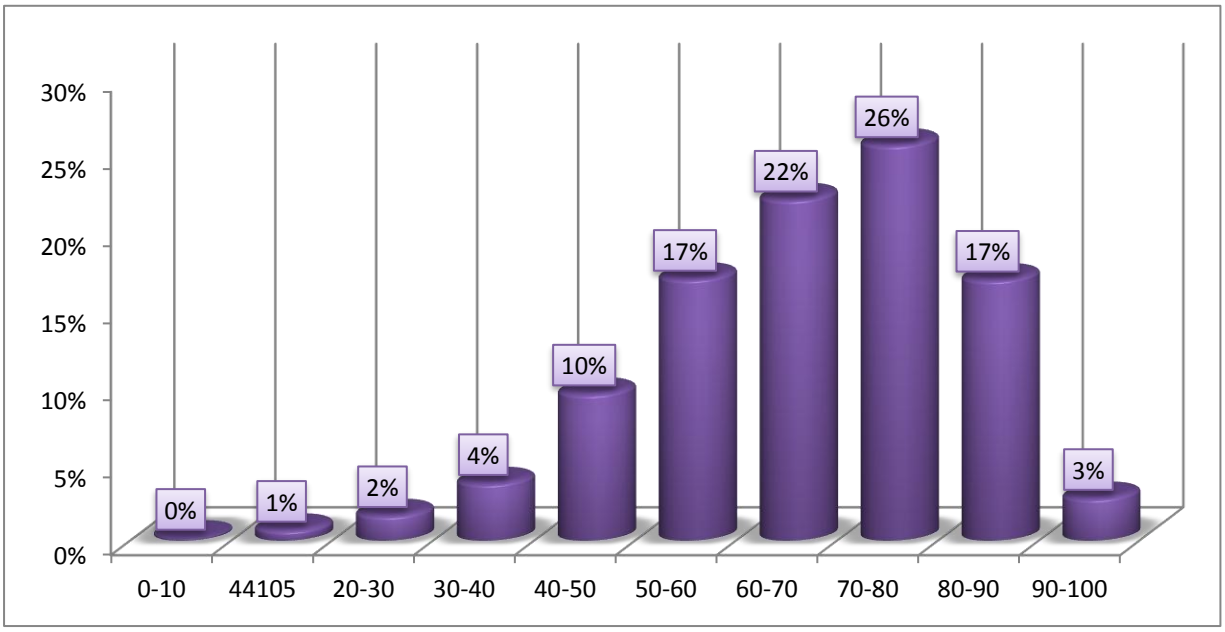
جدول (4–4) لفئات العمرية للمرضي

فئات العمر	العدد	النسبة %
0-10	161	0
10-20	691	1
20-30	1657	2
30-40	3775	4
40-50	9685	10
50-60	17256	17
60-70	22483	22
70-80	26068	26
80-90	17197	17
90-100	2793	3
المجموع	101766	100%

المصدر : اعداد الباحث باستخدام برنامج MS-Excel

الجدول 10 يوضح توزيع اعداد المرضي حسب الفئات العمرية المحددة ,نجد انه كلما كبر الانسان زادت نسبة اصابته بالسكري والشكل 3 يوضح أن العدد يتمركز في النصف الأكبر للاعمار.

الشكل ال(4–3) فئات العمرية للمرضي



المصدر : اعداد الباحث باستخدام برنامج MS-Excel

رابعاً : مدة الاقامة في المستشفي:–

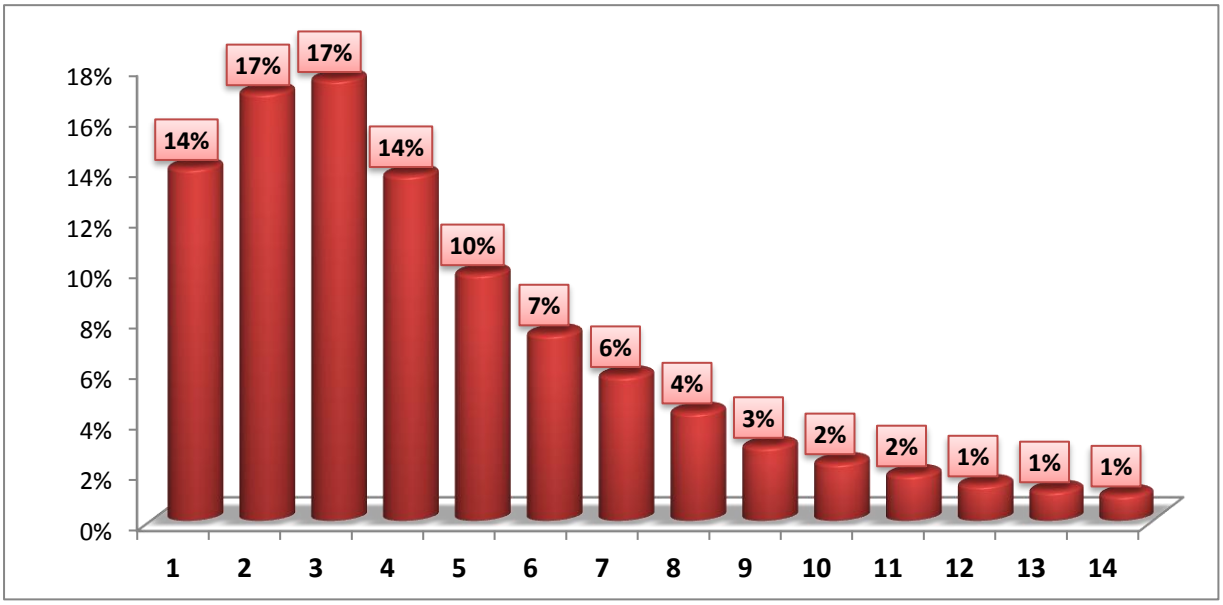
جدول (4–5) توزيع المرضي حسب عدد ايام الاقامة في المستشفي

عدد الايام	العدد	النسبة
1	14208	14%
2	17224	17%
3	17756	17%
4	13924	14%
5	9966	10%
6	7539	7%
7	5859	6%
8	4391	4%
9	3002	3%
10	2342	2%
11	1855	2%
12	1448	1%
13	1210	1%
14	1042	1%
المجموع	101766	100%

المصدر : اعداد الباحث باستخدام برنامج MS-Excel

الجدول(4–5) يوضح مدة اقامة المريض في المستشفي حيث بلغت اقصي مدة هي اسبوعين وأقلها يوم واحد ومجموع نسب أول ثلاثة ايام يساول (49%) مما يعني ان مرضي السكري عادة لا يقيمون مدة طويلة في المستشفي.

الشكل(4–4) توزيع المرضي حسب عدد ايام الاقامة في المستشفي



المصدر : اعداد الباحث باستخدام برنامج MS-Excel

خامسا : الاجراءات المعملية:

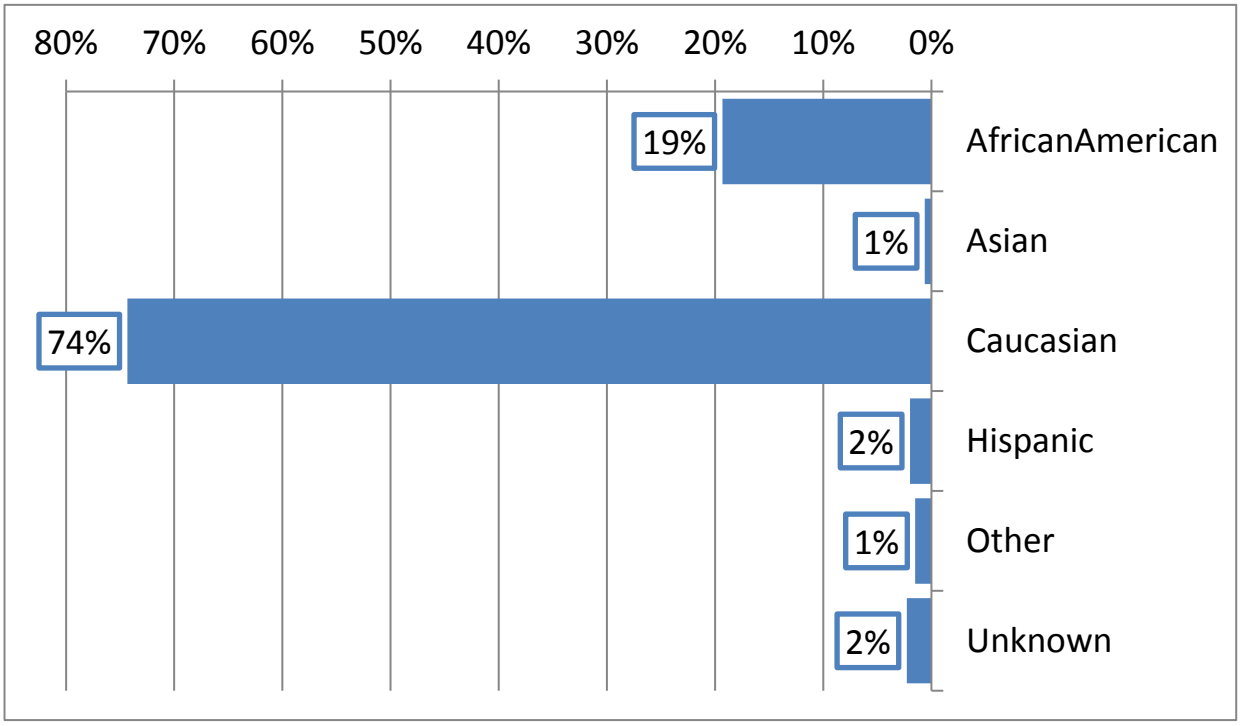
جدول (4–6) الاجراءات المعلمية

السلالة	العدد	متوسط عدد الاجراءات
امريكيون من اصول افريقيةAfrican American	846,874	44
آسيويونAsian	26,417	41
قوقازييونCaucasian	3,259,544	43
من اصل اسبانيHispanic	87,170	43
من اصول اخري او مختلطةOther	65,416	43
غير معلومي الاصلUnknown	100,250	44
المجموع	4,385,671	43

المصدر : اعداد الباحث باستخدام برنامج MS-Excel

يحتوي الجدول (4–6) علي متوسط عدد الاجراءات المعلملية التي تلقتها الحالة داخل المستشفي حيث بلغ المتوسط العام للإجراءت المعملية (43) إجراء ، في كل مقابلة ، و يمكن ملاحظة ان الاجراءات المعملية لسلالة القوقازييون بلغت (3,259,544) ، والتي تمثل (2%) من عدد المصابين في حزمة البيانات في الجدول (4–2).

الشكل (4–5) الاجراءات المعملية



المصدر : اعداد الباحث باستخدام برنامج MS-Excel

سادساً : الادوية أوالوصفات الطبية

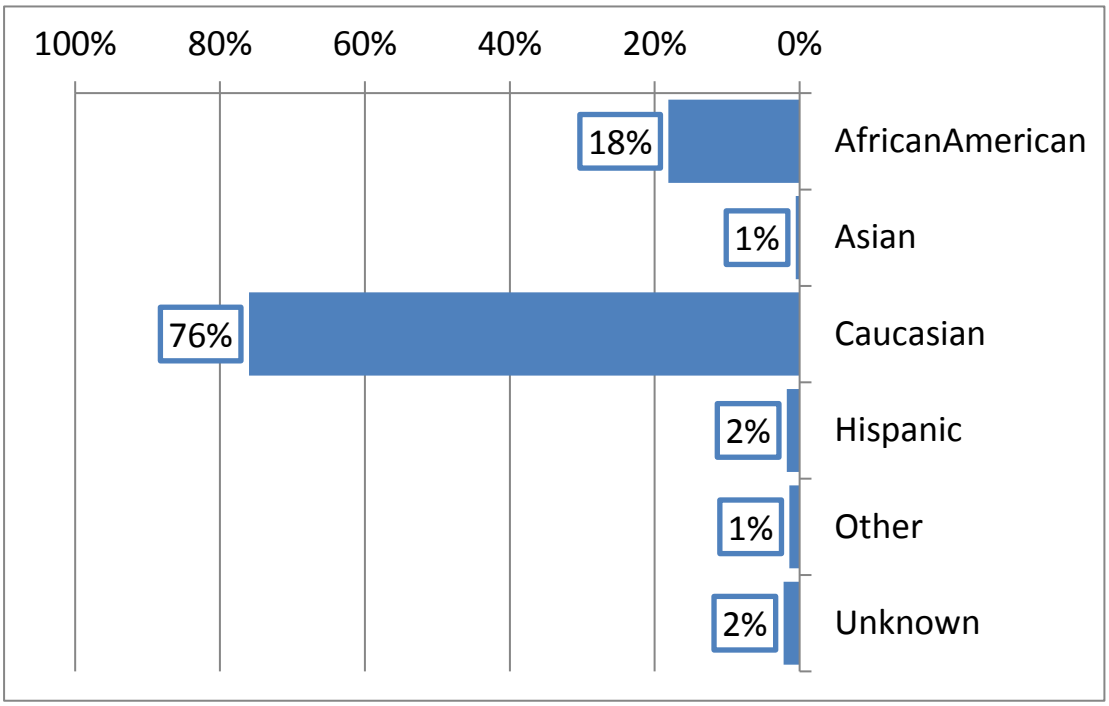
جدول (4–7) الادوية أو الوصفات الطبية

السلالة	المتوسط	العدد	النسبة
امريكيون من اصول افريقيةAfrican American	15	295,237	18%
آسيويونAsian	13	8,520	1%
قوقازييونCaucasian	16	1,239,328	76%
من اصل اسبانيHispanic	14	28,580	2%
من اصول اخري او مختلطةOther	15	22,870	1%
غير معلومي الاصلUnknown	15	35,944	2%
المجموع	16	1,630,479	100%

المصدر : اعداد الباحث باستخدام برنامج MS-Excel

يحتوي الجدول (4–7) علي متوسط عدد الادوية التي تلقتها كل حالة في حزمة البيانات حيث نلاحظ أن عدد الادوية الذي تلقتها سلالة القوقازييون بلغت (1,239,328) بنسبة (76%) وهو عدد منطقي مقارنة بعدد الإجراءات المعملية التي تتلقاها هذه السلالة في الجدول رقم (4–6).

الشكل (4–6) نسبة الادوية أو الوصفات العلاجية



المصدر : اعداد الباحث باستخدام برنامج MS-Excel

سابعاً : الإجراءات (غير المعملية)

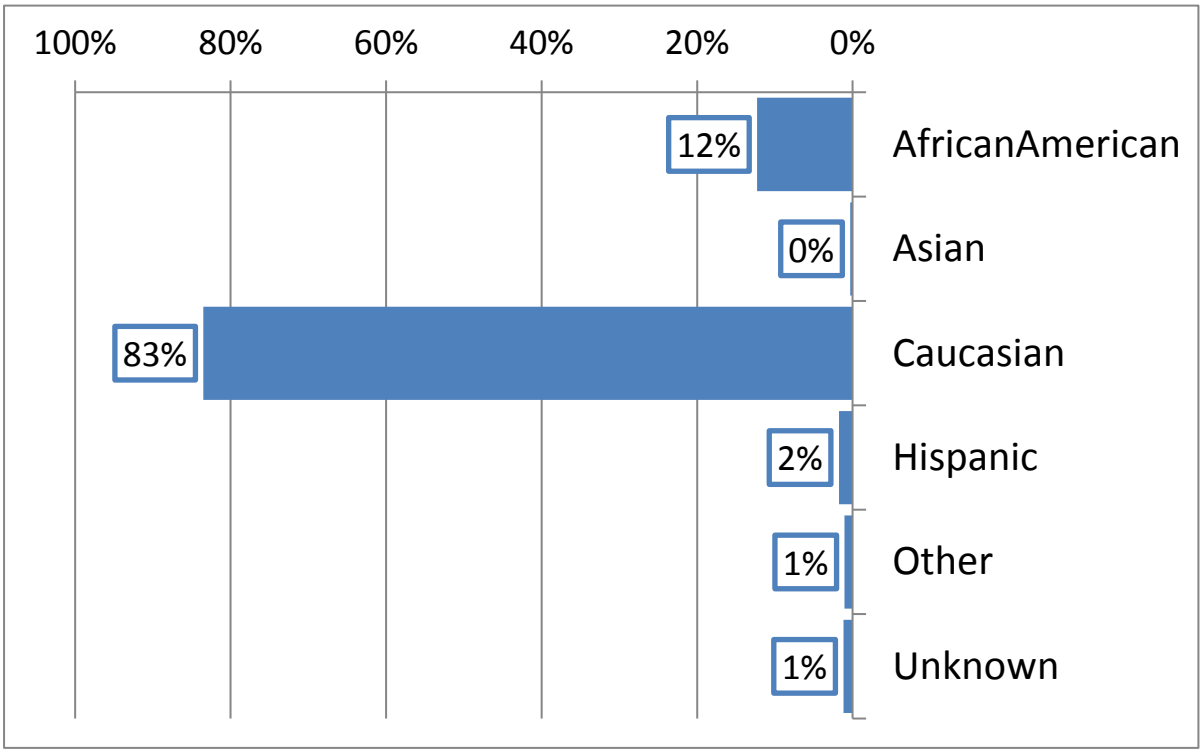
جدول (4–8)

	العدد	النسبة
امريكيون من اصول افريقيةAfrican American	4,609	12%
آسيويونAsian	106	0%
قوقازييونCaucasian	31,384	83%
من اصل اسبانيHispanic	661	2%
من اصول اخري او مختلطةOther	386	1%
غير معلومي الاصلUnknown	442	1%
المجموع	37,588	100%

المصدر : اعداد الباحث باستخدام برنامج MS-Excel

يحتوي الجدول (4–8) علي عدد الاجراءات (غير المعلمية) التي أجرتها الحالات في المستشفيات ونلاحظ ان عدد الاجراءات التي اجراها القوقازييون (31,384) بنسبة (83%) ، يليهم الامريكيون من اصول افريقية بنسبة بلغت (12%).

الشكل (4–7) نسبة الاجراءات (غير المعلمية)



المصدر : اعداد الباحث باستخدام برنامج MS-Excel

ثامناً : الزيارات الطارئة

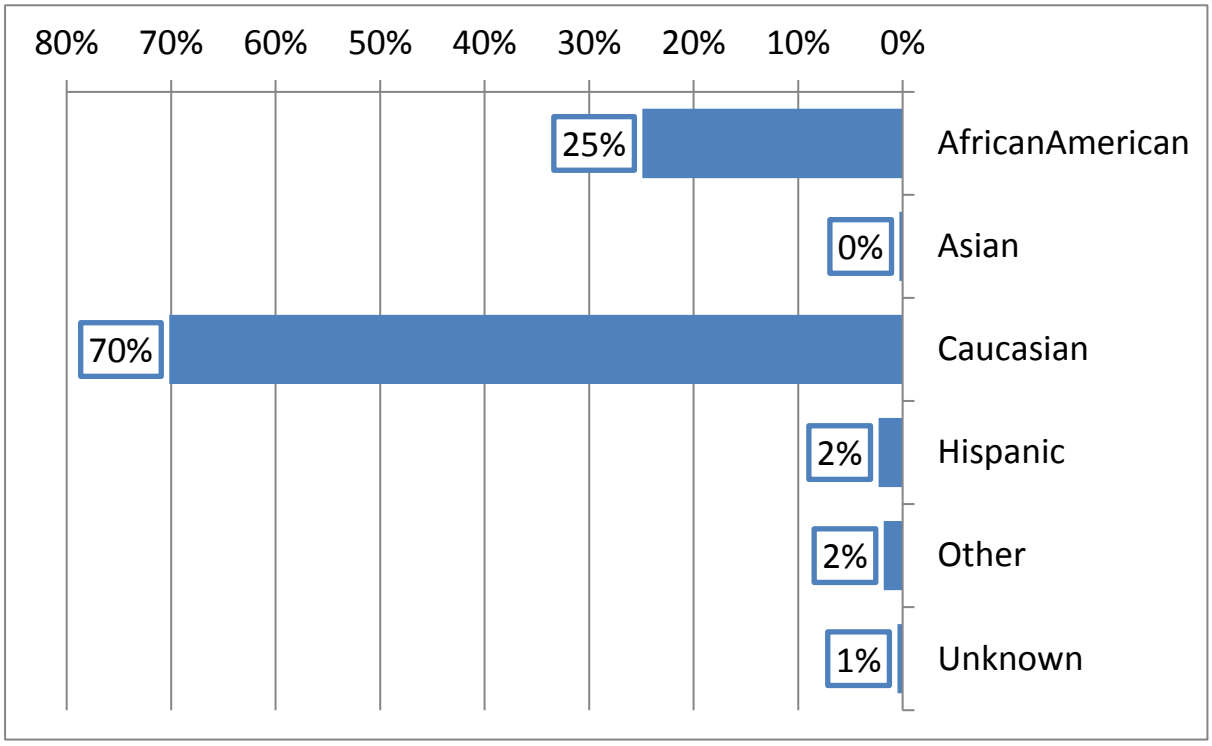
جدول (4–9) عدد الزيارات الطارئة

Row Labels	العدد	النسبة
امريكيون من اصول افريقيةAfrican American	5,014	25%
آسيويونAsian	60	0%
قوقازييونCaucasian	14,130	70%
من اصل اسبانيHispanic	465	2%
من اصول اخري او مختلطةOther	363	2%
غير معلومي الاصلUnknown	101	1%
Grand Total	20,133	100%

المصدر : اعداد الباحث باستخدام برنامج MS-Excel

الجدول (4–9) يوضح عدد الزيارت الطارئة التي تقوم بها الحالات خلال عام كامل ونلاحظ ايضاً أن القوقازييون هم أكثر السلالات التي قامت بزيارات طارئة الي المستشفيات بإجمال عدد زيارات (14,130) بنسبة بلغت (70%)

الشكل (4-8) نسبة الزيارات الطارئة



المصدر : اعداد الباحث باستخدام برنامج MS-Excel

4-4 تحليل البيانات:

فيما يلي خطوات كل من طريقتي المتوسطات (k-means) والهرمية (Hierarchical) حيث نعرض خطوات عمل كل طريقة في برنامج R مع تفسير مهمة كل الدول والمتغيرات الوارد ذكرها وتحديد مهمة محتوياتها ابتداءاً من اختيار العينات وصولا الي اختبار قياسات الزمن.

4-4-1 اختيار العينات

لاخذ عينات عشوائية من حزمة البيانات المستخدمة نتبع الخطوات الاتية والتي تعرض كيفية اختيار ثلاثة عينات عشوائية كمثال وتخزينها بالاسماء Sample1 ، Sample2 ، Sample3

1. Poplation.size <- sample(3,nrow(iris_data),replace = TRUE,prob = c(0.9,0.7,0.3))

في هذه الخطوة يقوم البرنامج بتخزين كل متغيرات الدراسة الموضوعة في ملف يسمي (iris_data) حيث تم تحديد حجم كل عينة من خلال تعريف نسبتها من الكل بواسطة (prob=) حيث نجد ان احجام العينات الاولي والثانية والثالثة، معرفة في المتجه (c(0.9,0.7,0.3)) علي التوالي

2. sample = Poplation.size
3. sample1<-iris_data[sample1==1,]

في الخطوتين 2 و 3 تم تخزين البيانات التي تم اختيارها عشوائيا من حزمة البيانات (iris_data) في العينة الاولي في المتغير Sample1 حيث يعبر الرقم (1==) عن العينة الاولي التي تم اخذها عشوائيا والرقم (,]) يعني انه سناخذ كل المتغيرات (الاعمدة Rows) الموجودة في حزمة البيانات بعين الاعتبار عند السحب العشوائي.

4. sample2<- iris_data[sample1==2,]

اختيار العينة الثانية

5. sample3<- iris_data[sample1==3,]

اختيار العينة الثالثة .

وهكذا يتم اختيار كل العينات الي ان نصل لحجم عينة يعجز فيها برنامج R عن اجراء طرق التحليل العنقودي باستخدام احدي الطريقتين نسبة لكبر حجم البيانات.

نعلم مما سبق ان حجم حزمة البيانات مكونة من (101,766) حالة وتم قياس (9) متغيرات موضحة في جدول (4–1) متغيرات الدراسة .

تم اختيار حجم العينة الاولي (1015) حالة " من تحديد الباحث" ، وتم اضافة نفس العدد للعينة الثانية عشوائيا ليصبح حجمها مضاعفاً (2030) وهكذا ، حيث تم الحصول علي (100) عينة عشوائية لإخضاعها لتجربة قياس الزمن المستغرق لاعطاء نتائج التحليل وهو العامل الرئيسي الذي تم استخدامه لتقييم طريقتي التحليل العنقودي بالاضافة الي مناقشة اداء كل طريقة حسب قياسات الزمن التي تمت لها ، تم الاكتفاء بقياس الزمن المستغرق لاول (30) عينة تم الحصول عليها لاقتراب هذا العدد من التوزيع الطبيعي عند اجراء الاختبارات الاحصائية.

4–4–2 تطبيق الطريقتين

تم تطبيق كل طريقة بصورة منفصلة ، لنفس العينات المأخوذة عشوائيا لضمان حساب الزمن بدقة مع مراعات ان العينة المأخوذة مخزنة مسبقا في ملف MS-Excel منفصل يتم تحميله الي البرنامج اولا قبل البدء في حساب الزمن بواسطة الدالة ptm ولضمان الشفافية في اعطاء النتائج لكل طريقة .

تم التركيز علي حساب الزمن الذي تاخذه منهجية كل طريقة لحساب النتائج وليس لعرضها علي شاشة الكمبيوتر ، احيانا عرض الرسومات البيانية والجداول الكبيرة ياخذ بعض الوقت من زمن الجهاز لعرضها بصورة اوضح مما يستدعي اخراج الدوال الخاصة بعرض النتائج والرسومات البيانية خارج دالة الزمن لتكون عملية حساب الزمن فقط لدالة كل طريقة لوحدها دون حصول اي زيادة للزمن بسبب عملية اخري.

عند محاولة اختبار نفس الخطوات السابقة لكلا الطريقتين يجب مراعاة استخدام جهاز بنفس المواصفات[6] و ضرورة الحصول علي اصدار مطابق لنفس رقم الاصدار[7] المستخدم لبرنامج R في هذا البحث لضمان الحصول علي نفس النتائج او قريبة منها بسبب ان اختيار العينات يتم عشوائيا.

يجب ملاحظة انه تم التركيز علي خطوات العمل الفعلي لكلا الطريقتين مع تجاهل النتائج النهائية الخاصة بمرضي السكري بسبب تناولها في بحوث كثيرة متعمقة ومتخصصة أكثر ونحن هنا بصدد دراسة وتقييم اداء الطريقتين حيث تم اعتماد الزمن كمعيار رئيسي ومهم لتقييم الطريقتين.

تم الحصول علي الزمن بالثواني مع تغيير حجم العينة واخذها عشوائيا في كل مرة كما هو موضح حيث تم أخذ (30) عينة عشوائية من مجموعة البيانات (Data Set) المذكورة سابقا[8] .

تمت مقارنة الطريقتين باستخدام المتغير "Elapsed Time" باعتباره الزمن الكلي منذ بداية تنفيذ الأمر وحتي نهايته بالحصول علي النتائج داخل الذاكرة الناتج من استخدام الدالة

proc.time ()

حيث تقوم هذه الدالة بتخزين التوقيت في متغير قبل بداية التنفيذ مباشرة ومن ثم تحسب الزمن بعد التنفيذ ويتم طرح القيمة الاولي منه للحصول علي الزمن المستغرق حيث يظهر النتاج في شكل ثلاثة قيم كما هو موضح في الخطوات الاتية[9]:-

```
Start.point <- proc.time()
“Type your code here يجب كتابة كود البرمجة في هذه المساحة ”
proc.time () – Start.point
Start.point ≡ “the time on Start”.
Proc.time ()≡ “function retrieve the time from Windows” .
```

تنبيه:(يجب تظليل الثلاثة خطوات عند التنفيذ حتي يقوم برنامج (R) بتنفيذ الخطوات بالتتابع).

[6] Compaq Presario CQ61, Ram (2GB), Processor Pentium Dual-Core CPU 2.20GHz, System Windows 7 Ultimate 64Bit.

[7] برنامج (R-Programming) (النسخة R i386 3.3.2)

[8] مرجع سابق رقم (4)

[9] Documentation for R Base Package ‘version 3.3.2

4-4-3 خطوات عمل طريقة المتوسطات K-means في برنامج R :

1. kmeans_data = read.csv("E:/my phd/Third chapter/Random Sample for Analysis/sample100.csv")

الخطوة الاولي ادخال الملف الذي يحتوي علي البيانات الي برنامج R حيث يستقبل البرنامج انواع مختلفة من الملفات وهنا نستخدم الملف ذو الصيغة (.CSV) والتي تعني (Comma Separated Value) القيم المعرفة بفاصلة اي ان الفاصلة (,) هي الرمز الفاصل بين كل قيمة والاخري وهذه الصيغة توفر في المساحة الكلية للملف وتستغل مساحة صغيرة من ذاكرة الحاسوب.

2. kmeans_data_features <- kmeans_data

تخزين جدول البيانات في متغير

3. ptm <- proc.time()

وضع قيمة الزمن الحالية في متغير يسمي ptm , يحيث يقوم البرنامج بوضع الزمن الحالي في لنظام الحاسوب في هذا المتغير

4. result <- kmeans(kmeans_data_features,4)

الدالة الرئيسية لحساب العناقيد بطريقة المتوسطات هي ()Kmeans والمتغير result هو من تسمية الباحث لتخزين النتائج فيه بدلا عن عرضها في شاشة البرنامج والعدد (4) يمثل القيمة (K) التي تحتاجها طريقة المتوسطات لحساب عدد العناقيد بوضع نقاط عشوائية وسط الشكل الانتشاري للبيانات بعدد (k) ومنها تبدأ مرحلة تجميع كل العناصر الي العناقيد الاقرب لها ويتم اعادة حساب النقاط لتكون مركزية (في الوسط تمام) ويعاد حساب المسافات بين هذه النقاط وبقية العناصر مرة اخري لمتحديث انتماء العناصر الي العناقيد ويتم تكرار هذه الخطوة داخليا في البرنامج حتي تصل النقاط المركزية في حالة سكون (تكون في المركز تماماً حيث تكون الآخيره هذه هي العناقيد المطلوبة .

5. proc.time() - ptm

تقوم هذه الدالة بطرح الزمن الحالي من الزمن الذي تم وضعه في المتغير ptm سابقا لحساب الزمن الذي استغرقه البرنامج لاعطاء نتائج طريقة المتوسطات في برنامج R

6. result

هذا المتغير يحتوي علي النتائج التي تم الحصول عليها في الخطوة (4) ويقوم بعرضها علي شاشة البرنامج في شكلها النهائي

7. cluster_member <- table(kmeans_data$age, result$cluster)

في هذه الخطوة يقوم البرنامج بتخزين جدول يحوي علي تصنيف متغير اعمار المرضي في كل عنقود في متغير يسمي cluster_member

8. cluster_member

لعرض الجدول في الخطوة (7) يجب كتابة اسم المتغير الذي يحتوي علي جدول عرض بيانات العناقيد حسب متغير عمر المريض

9. plot(kmeans_data$petal.length,kmeans_data$petal.width , col=result$cluster)
10. plot(kmeans_data$petal.length,kmeans_data$petal.width , col=kmeans_data$class)
11. plot(kmeans_data$sepal.length,kmeans_data$sepal.width , col=result$cluster)
12. plot(kmeans_data$sepal.length,kmeans_data$sepal.width , col=kmeans_data$class)

الخطوات الاربعة الاخيرة لعرض رسومات بيانية تخص طريقة المتوسطات في التحليل العنقودي

13. Population.size <- sample(3,nrow(iris_data),replace = TRUE,prob = c(0.9,0.7,0.3))

هذه الخطوة توضيح كيفية اخذ عينات عشوائية من حزمة البيانات المستخدمة في التحليل حيث تعمل ادالة Sample علي اختيار العينات العشوائية .

جدول (4–10) قياسات الزمن لطريقة المتوسطات K-means:

رقم العينة	حجم العينة n	زمن المستخدمUser	زمن النظام System	الزمن الكلي Elapsed
1	1015	0	0	0.05
2	2030	0.02	0	0.01
3	3045	0.01	0.01	0.03
4	4060	0	0	0.02
5	5075	0.01	0	0.01
6	6090	0.01	0	0.01
7	7105	0.02	0	0.02
8	8120	0.02	0	0.02
9	9135	0.01	0	0.05
10	10150	0.05	0	0.34
11	11165	0.06	0.01	1.09
12	12180	0.03	0	1.29
13	13195	0.06	0	0.64
14	14210	0.06	0	0.89
15	15225	0.06	0.02	5.03
16	16240	0.14	0.01	0.85
17	17255	0.06	0	0.64
18	18270	0.03	0	0.3
19	19285	0.08	0.02	2.24
20	20300	0.13	0	0.73
21	21315	0.07	0.01	0.5
22	22330	0.13	0.02	0.82
23	23345	0.04	0	0.38
24	24360	0.06	0.03	0.28
25	25375	0.06	0	0.35
26	26390	0.28	0.2	0.98
27	27405	0.1	0	0.31
28	28420	0.09	0.01	0.23
29	29435	0.1	0.02	0.22
30	30450	0.24	0.02	5.77

المصدر : اعداد الباحث باستخدام برنامج R

User: قياس الزمن الذي يستغرقه المستخدم للعملية الحالية.

System: قياس الزمن الذي استغرقه النظام لتنفيذ العملية الحالية.

Elapsed : قياس الزمن الكلي منذ بداية تنفيذ العملية وحتي نهايتها.

يحتوي الجدول (4-10) علي قياسات الزمن لطريقة المتوسطات وحجم كل عينة تم اخذها وقياسات الزمن بالتفصيل لكل من المستخدم والنظام والزمن الذي استغرقته طريقة المتوسطات في تنفيذ كافة خطواتها واعطاء النتائج النهائية ، فنجد مثلا في العينة الاولي استغرقت طريقة المتوسطات (0.05) من الثانية لاعطاء نتائج العينة الأولي ، و (0.01) لاعطاء نتائج العينة الثانية وهكذا ، ويرجع الاختلاف في الزمن الي حجم العينة المختارة عشوائياً فكل عينة تختلف في مكوناتها وحجم الارقام بداخلها عن الاخري.

كما نلاحظ ان اكبر قيمة للزمن كانت تخص العينة رقم (30) والتي شملت علي (30,450) حالة حيث استغرقت (5.77) ثانية ، لنجد انه زمن مقبول جدا في اطار البيانات الكبيرة مما يتيح لمتخذ القرار الحصول علي التوصيات المستخرجة من تحليل هذه الكمية من البيانات في وقت مناسب ليتمكن من اتخاذ قراره ، واذا نظرنا اليها من وجهة نظر الأعمال ، فيمكن ان نقول ان صاحب العمل يستطيع اغتنام فرص الاستثمار في وقتها وقبل فواتها.

4- 4-4 خطوات عمل طريقة الهرمية Hierarchical في برنامج R :

1. hc = read.csv("E:/my phd/Third chapter/Random Sample for Analysis/sample13.csv")

الخطوة الاولي ادخال الملف الذي يحتوي علي البيانات الي برنامج R حيث يستقبل البرنامج انواع مختلفة من الملفات وهنا نستخدم الملف ذو الصيغة (.CSV) والتي تعني (Comma Separated Value) القيم المعرفة بفاصلة اي ان الفاصلة (,) هي الرمز الفاصل بين كل قيمة والاخري وهذه الصيغة توفر في المساحة الكلية للملف وتستغل مساحة صغيرة من ذاكرة الحاسوب.

2. hc.cluser = hc

تخزين البيانات في متغير يسمي hc.clust لضمان الاحتفاظ بنسخة من البيانات الاصلية قبل اجراء اي تعديلات عليها.

3. ptm <- proc.time()

وضع قيمة الزمن الحالية في متغير يسمي ptm , يحيث يقوم البرنامج بتخزين قيمة الزمن الحالي لنظام الحاسوب في هذا المتغير

4. hc.method <- hclust(dist(hc.cluser), method = "complete")

في هذه الخطوة يقوم البرنامج بحساب العناقيد بواسطة الطريقة الهرمية باستخدام الدالة (hclust) وتخزينها في متغير يسمي hc.method ، حيث يحتوي هذا المتغير علي كافة مخرجات الطريقة من جداول

ومصفوفات وصولا الي عدد العناقيد الذي تم التوصل اليه وخطوات انتماء كل العناصر الي العناقيد والتي توضح في الشكل رقم (3-4).

5. proc.time() – ptm

تقوم هذه الدالة بطرح الزمن الحالي للجهاز الحاسوب من الزمن الذي تم وضعه في المتغير ptm سابقا لحساب الزمن الذي استغرقه البرنامج لاعطاء نتائج طريقة المتوسطات في برنامج R

6. hc.method

عرض النتائج المخزنة في المتغير hc.metho حيث يعرض كل النتائج المتحصل عليها من طريقة الهرمية.

7. plot(hc.method,main = "Hierarchical Clustering Test 1" ,cex=0.9)

8. plot(hc.cluser$sepal.length,hc.cluser$sepal.width,col = hc$class)

الخطوتين 7و8 لعرض بعض الرسومات البيانية الخاصة بطريقة الهرمية.

جدول (4–11) قياس الزمن للطريقة الهرمية Hierarchical Analysis

رقم العينة	n حجم العينة	زمن المستخدمUser	زمن النظام System	الزمن الكليElapsed
1	1015	0.11	0	0.14
2	2030	0.44	0.03	0.47
3	3045	1.08	0.11	1.23
4	4060	2.06	0.03	2.29
5	5075	3.09	0.12	3.23
6	6090	4.46	0.13	4.72
7	7105	5.95	0.34	6.44
8	8120	7.88	0.52	8.56
9	9135	9.99	1.57	17.07
10	10150	12.6	1.01	20.53
11	11165	15.6	1.8	83.81
12	12180	18.5	2.47	211.67
13	13195	Error	Error	Error
14	14210	Error	Error	Error
15	15225	Error	Error	Error
16	16240	Error	Error	Error
17	17255	Error	Error	Error
18	18270	Error	Error	Error
19	19285	Error	Error	Error
20	20300	Error	Error	Error
21	21315	Error	Error	Error
22	22330	Error	Error	Error
23	23345	Error	Error	Error
24	24360	Error	Error	Error
25	25375	Error	Error	Error
26	26390	Error	Error	Error
27	27405	Error	Error	Error
28	28420	Error	Error	Error
29	29435	Error	Error	Error
30	30450	Error	Error	Error

المصدر : اعداد الباحث باستخدام برنامج R

User: قياس الزمن الذي يستغرقه المستخدم للعملية الحالية.

System: قياس الزمن الذي استغرقه النظام لتنفيذ العملية الحالية.

Elapsed : قياس الزمن الكلي منذ بداية تنفيذ العملية وحتي نهايتها.

Error : عدم مقدرة البرنامج علي اعطاء النتائج او المخرجات الخاصة بالهرمية.

تم اختيار اول (30) عينة لتجربة قياس الزمن لإقترابها من التوزيع الطبيعي حيث نلاحظ ان العينة الاولي شملت (1015) حالة مختارة عشوائيا من حزمة البيانات ونلاحظ ازدياد الزمن في كل عينة حيث استغرقت العينة الاولي (0.14) ثانية لاعطاء النتائج النهائية لعناقيد الطريقة الهرمية واستغرقت العينة الثانية (0.47) ثانية ، وفي كل مرة يتم فيها اخذ عينة جديدة يتم مضاعفة الحجم لمعرفة تاثير حجم العينة علي الزمن الذي تستغرقه الطريقة الهرمية لاعطاء النتائج النهائية ونلاحظ توقف البرنامج في العينة رقم (13) واعطاء خطأ بعدم مقدرة النظام علي اعطاء النتائج او المخرجات الخاصة بالتحليل نسبة لان حجم المخرجات اكبر من 664.1 ميقا بايت او اكثر وذلك يعني ان هذه الطريقة تحتاج الي مساحة اكبر في الذاكرة المؤقته وعجز البرنامج عن تنفيذ خطواتها .

اذا رجعنا الي الخطوات العملية التي تنفذها هذه الطريقة كما هو موضح في الفصل السابق نجد ان اكبر خطوة قد تأخذ مساحة في الذاكرة المؤقته موجود في الجدول رقم (3-1) وهي مصفوفة القرابة والتي تقوم بعمل مصفوفة مربع تكون ابعادها بعدد الحالات المدخلة الي النظام وفي العينة رقم (13) تم ادخال (13195) حالة ، عندها يعجز النظام عن استخدام اي من المسافات التي تقوم بحساب المسافة بين كل حالة والاخري فمنهجية هذه الطريقة تجعل من المستحيل استخدامها في البيانات الكبيرة .

4-4-5 اختبار توزيع قياسات الزمن

بعد الحصول علي متغير الزمن وجب علينا اختبار توزيع قيم الزمن المستخدمة لمعرفة هل تتبع التوزيع الطبيعي أم لا.

الجدول يوضح إختبار (Normality Test) واخذ قيمة الـــ Sig من إختبار(Shapiro-Wilk) وكانت النتائج كاللآتي:-

فيما يلي ملخص يوضح اسم المتغير وعدد القيم في كل طريقة ونسبتها وكذلك القيم المفقودة.

الجدول (4–12) اسم المتغير وعدد القيم في كل طريقة

ملخص الحالات التي تمت معالجتها						
اسم المتغير	الطريقة	الحالات				
		القيم		القيم المفقودة		المجموع
		العدد	النسبة	العدد	النسبة	العدد	النسبة
الزمن الكلي للتنفيذ	المتوسطات	30	100%	0	0.0%	30	100%
	الهرمية	12	100%	0	0.0%	12	100%

المصدر : إعداد الباحث باستخدام برنامج SPSS 20

احتوي الجدول (4–12) علي وصف لعدد العينات التي تم قياس الزمن لها حيث احتوت الطريقة الهرمية علي (12) عينة وتم توضح سبب حصولنا علي هذا العدد في (4–7) ، احتوت طريقة المتوسطات علي (30) عينة ، يمكن ملاحظة ان القيم المفقودة كانت بنسبة (0%).

4–4–6 اختبار الانتماء للتوزيع الطبيعي

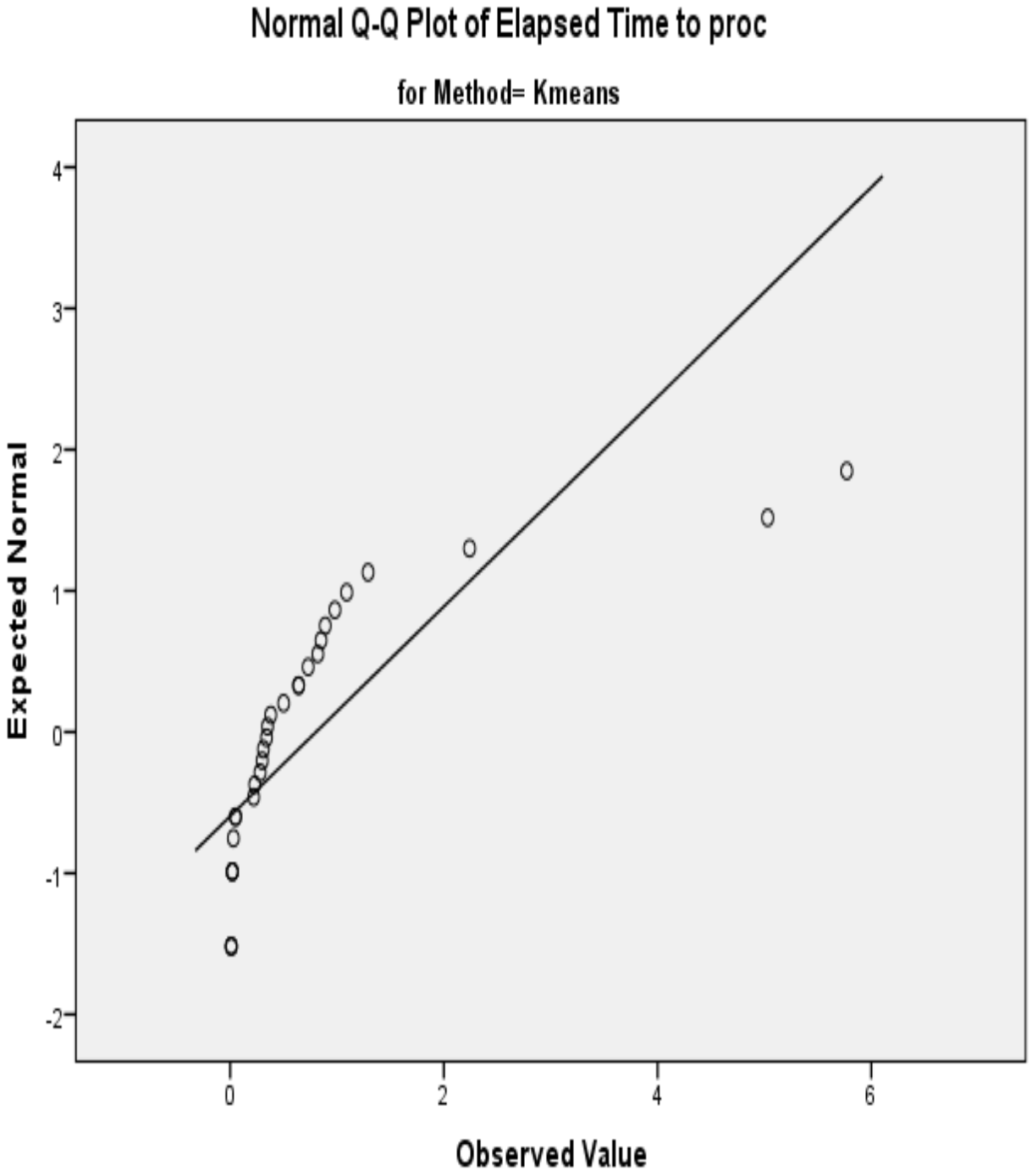
الجدول (4–13)

اسم المتغير	الطريقة	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
		قيمة الاختبار	درجة الحرية	القيمة المعنوية	قيمة الاختبار	درجة الحرية	القيمة المعنوية
الزمن الكلي للتنفيذ	المتوسطات	0.282	30	0	0.576	30	0
	الهرمية	0.394	12	0	0.544	12	0
a. Lilliefors Significance Correction							

المصدر : إعداد الباحث باستخدام برنامج SPSS 20

الجدول (4–13) يوضح قيمة اختبار التوزيع الطبيعي وفيه نقارن نسبة الـــSig في اختبار (Shapiro-Wilk) وهي (0) بقيمة (0.05) لنجد ان قياسات الزمن بالثواني لا تتوزع طبيعيا مما يجعلنا نستخدم اختبار (T) اللامعلمي لاختبار الفرق بين متوسطي عينتين مستقلتين (two independent samples test) ، لمعرفة هل توجد فروق معنوية متوسط الزمن المستغرق لكل طريقة .

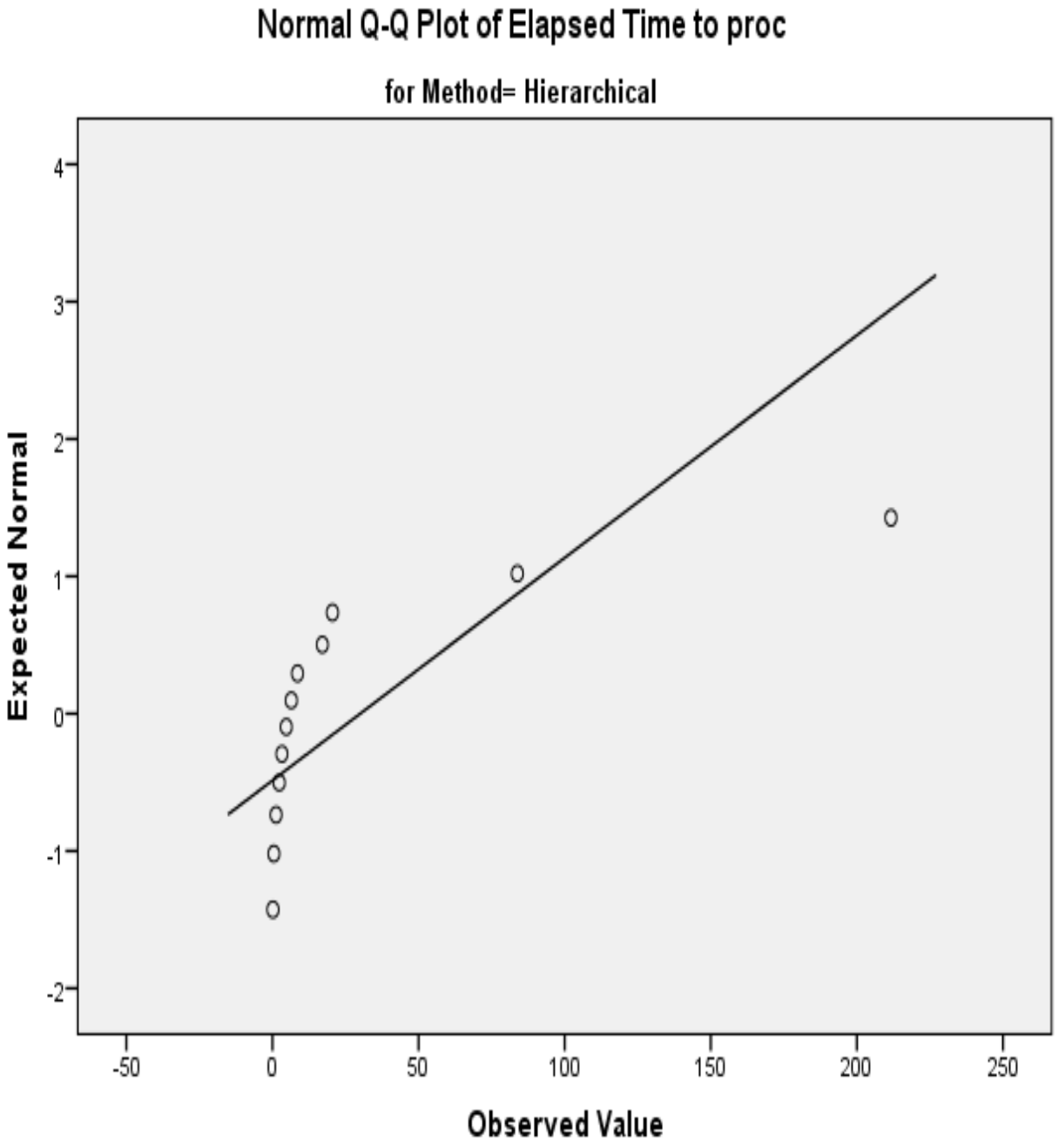
شكل (4–9) ملائمة البيانات للتوزيع الطبيعي في طريقة المتوسطات باستخدام (Normal Q-Q Plot)



المصدر : إعداد الباحث باستخدام برنامج SPSS 20

الشكل (4–9) يوضح تشتت قيم متغير الزمن الكلي في طريقة ال K-means بجانب الخط المستقيم الذي يمثل التوزيع الطبيعي , فكلما كانت القيم مشتتة بجانب الخط المستقيم تكون ذات توزيع طبيعي اما اذا كانت مشتته خارج الخط المستقيم كما في الشكل 5 فهذا يدل علي عدم اتباعها التوزيع الطبيعي.

شكل (4–10) ملائمة البيانات للتوزيع الطبيعي في الطريقة الهرمية باستخدام (Normal Q-Q Plot)



المصدر : إعداد الباحث باستخدام برنامج SPSS 20

الشكل (4–10) يوضح تشتت قيم متغير الزمن في طريقة (Hierarchical clustering) مقارنة بالخط المستقيم الذي يمثل التوزيع الطبيعي ونجد انها مشتت بعيدا عن الخط المستقيم مما يدل علي عدم اتبعاها التوزيع الطبيعي.

4-4-7 إختبار الفروق المعنوية لقياسات الزمن في الطريقتين:

تم استخدام اختبار (Two-Sample Kolmogorov-Smirnov Test) اللا معلمي لعينتين مستقلتين[10] كنتيجة لعدم اتباع قياسات الزمن للتوزيع الطبيعي وظهرت النتائج الآتية :-

جدول (4-14) وصف لقيم كل متغير ومتوسطه وانحرافه المعياري.

المتغير	الطريقة	العدد	المتوسط	الانحراف المعياري	الخطا المعياري للمتوسط
زمن المستخدم	المتوسطات	30	0.0677	0.0654	0.012
	الهرمية	12	6.8133	6.1767	1.7831
زمن النظام	المتوسطات	30	0.0127	0.0365	0.0067
	الهرمية	12	0.6775	0.8393	0.2423
الزمن الكلي	المتوسطات	30	0.8033	1.3469	0.2459
	الهرمية	12	30.013	61.708	17.813

المصدر : إعداد الباحث باستخدام برنامج SPSS 20

احتوي الجدول (4-14) علي وصف لقياسات الزمن لكل من المستخدم والنظام والزمن الذي الذي استغرقته كل طريقة ، بلغ متوسط قياسات الزمن الكلي لطريقة المتوسطات (0.80) ثانية بانحراف معياري (1.3) ، ومتوسط قياسات الزمن للطريقة الهرمية (30.13) ثانية بانحراف معياري (61.7) مما يدل علي الاختلاف الكبيرة لقياسات الزمن عند زيادة حجم العينة في الطريقة الهرمية ونلاحظ الاختلاف الكبير بين الطريقتين في قياسات الزمن للمستخدم والنظام كذلك .

[10] IBM SPSS 20

جدول (4–15) اختبار Kolmogorov-Smirnov Z

الاختبار الاحصائي				
		زمن المستخدم	زمن النظام	الزمن الكلي
Most Extreme Differences	القيمة المطلقة	0.917	0.850	0.700
	ايجابي	0.917	0.850	0.700
	سلبي	0.000	0.000	0.000
Kolmogorov-Smirnov Z		2.684	2.489	2.049
Asymp. Sig. (2-tailed)		0.00	0.00	0.00
a. Grouping Variable: Method				

المصدر : اعداد الباحث باستخدام برنامج SPSS

الجدول (4–15) يوضح قيم اختبار (Two-Sample Kolmogorov-Smirnov Test)[11] والتي تم اجرائها لاختبار ما اذا كان هناك فروق ذات دلالة احصائية لقياسات الزمن لطريقتي الهرمية والمتوسطات.

وعليه ومن خلال ملاحظة نتائج الاختبار نجد ان هناك فروق ذات دلالة احصائية بين متوسط الزمن اللازم لتنفيذ طريقة المتوسطات والطريقة الهرمية حيث بلغت قيمة اختبار (Kolmogorov-Smirnov Z) لكل من زمن المستخدم والنظام والكلي (2.68) ، (2.48) ، (2.04) علي التوالي وبلغت قيمة اختبار المعنوية Sig (0) لكل قياس .

ومن خلال ملاحظة البيانات المتحصل عليها من برنامج R الموجودة في الجدول (4–10) والجدول (4–11) تعتبر طريقة المتوسطات أكفاء من حيث زمن التنفيذ واكثر ملائمة لحل مشكلة البيانات الكبيرة من الطريقة الهرمية لتميز منهجيتها باستغلال موارد الحاسب المتوفرة استغلال امثل واعتمادها علي وضع نقاط مركزية لتحديد انتماء اي مفردة للعنقود الاقرب لها من حيث المسافة .

[11] Smirnov, N. V. 1948. Table for estimating the goodness of fit of empirical distributions. Annals of the Mathematical Statistics, 19, 279-281.

فيما يلي جدول يوضح المقارنة بين طريقة المتوسطات والطريقة الهرمية من خلال عرض مقاييس الاختلاف بين الطريقتين ومدي ملائمة كل طريقة للبيانات الكبيرة

جدول (4-16) مقارنة معيارية بين طريقة الهرمية والمتوسطات

م	وجه المقارنة	الطريقة الهرمية	طريقة المتوسطات	ملاحظة
.1	عدد العينات المستخدمة في التحليل	12	30	توقفت الطريقة الهرمية عن اعطاء النتائج بعد العينة رقم 12 والتي احتوت علي 12180 حالة ، مقارنة بطريقة المتوسطات والتي استمرت حتي العينة رقم 30
.2	متوسط الزمن المستغرق لاعطاء النتائج	30.013 ثانية	0.8 ثانية	من خلال قراءة قيم المتوسط ومقارنتها بقيم الانحراف المعياري يتضح ان طريقة المتوسطات اكفأ من الهرمية في متوسط الزمن المستغرق لاعطاء النتائج
.3	الانحراف المعياري لمتوسط الزمن المستغرق لاعطاء النتائج	61.7	1.3	يتضح ان الانحراف المعياري لمتوسط الزمن لطريقة المتوسطات اقل من الهرمية
.4	الزمن المستغرق في اخر عينة تم اختبارها	211.67 ثانية	5.77 ثانية	الزمن المستغرق في طريقة الهرمية اكبر من طريقة المتوسطات
.5	مدي التاثر بزيادة حجم العينة	تتأثر	تأثير طفيف	
.6	وجود تفصيل واضح لمراحل تكوين العناقيد	يوجد	لايوجد	اهتمام طريقة الهرمية باعطاء تفصيل لمراحل تكوين العناقيد هو ما يجعلها تستغرق مزيد من الزمن في حالة العينات الكبيرة مما يحسب عليها

م	وجه المقارنة	الطريقة الهرمية	طريقة المتوسطات	ملاحظة
.7	امكانية العمل في العينات الصغيرة	تعمل	تعمل	كلا الطريقتين تعمل بكفاءة في العينات الصغيرة
.8	امكانية العمل في العينات الكبيرة	لا تعمل	تعمل	توقف طريقة الهرمية عن العمل علي عدد 13195 حالة لا يؤهلها للعمل علي العينات الكبيرة
.9	توزيع الزمن المستغرق لاعطاء النتائج	لا يتوزع طبيعيا	لا يتوزع طبيعيا	تم اخضاع الطريقتين لاختبار التوزيع الطبيعي