1-0: تمهيد

اصبح عالمنا اليوم مليء بالبيانات بمختلف انواعها من صور وفيديوهات ونصوص وارقام وباختلاف انواع البيانات تختلف انواع الملفات التي تحوي هذه البيانات والبرامج التي تتعامل معها فقد ظهر مصطلح البيانات الضخمة (Big Data) وصار من هواجس هذا العصر بالنسبة للمؤسسات الكبيرة والشركات العملاقة من عدة نواحي:

1. تخزين البيانات.
2. سهولة استرجاعها
3. التحليل الاحصائي وهو أهم خطوة لهذه المؤسسات.

يطلق علي أي بيانات مسمي "بيانات ضخمة" اذا توفرت فيها اغلب هذه الشروط:

1. السرعة العالية : تعني ان البيانات تتحرك بسرعة عالية وتتغير من مكان الي اخر فإذا لم تستفد منها اليوم أتت بيانات غيرها غدا بنفس الحجم او اكبر كمثال قد يستقبل محرك البحث قوقل "Google" 4 ملايين نص يراد البحث عنه في الدقيقة الواحدة .
2. الحجم الكبير: تعني أننا نتعامل مع بيانات بحجم البيتا بايت
   Peta Byte = 1,024 Terabytes / Terabyte=1,024 Gigabytes .
3. الاختلاف : تعني ان البيانات تحتوي علي انواع مختلفة غير البيانات المعتادة مثل الصور ومقاطع الصوت والفيديوهات وسجلات الحاسوب .
4. القيمة : تعني أن البيانات تحتوي علي قيمة عالية من المعلومات والتي يمكن الاستفادة منها فاذا لم نستطع ان نستخرج قيمتها عبر التحليل اليوم سنفقد فرصة الاستفادة منها.
5. الصلاحية : تعني ان هذه البيانات تحتوي علي بيانات يمكن الاستفادة منها وبيانات اخري يمكن تجاهلها أو إخراجها من التحليل الاحصائي. [1]

من خلال ما سبق نخلص الي أن البيانات الضخمة هي بيانات متغيرة الشكل وتنتقل سريعا من مكان الي اخر عبر الشبكات وحجمها كبير جدا وفيها كم هائل من المعلومات بعضها ذو فائدة عظيمة والبعض الاخر غير مهم مؤقتاً.

مسألة تخزين البيانات واسترجاعها لا تدخل في سياغ هذه الدراسة ولكن ما يهم هو جانب التحليل الاحصائي لهذه البيانات , مع العلم أن محتوي هذا النوع من البيانات في معظم الأوقات غير معلوم مما يجعل أساليب الاستكشاف الاحصائية التي تعمل في هذا الجانب تأخذ حيزاً

[1] https://aci.info/2014/07/12/the-data-explosion-in-2014-minute-by-minute-infographic

مهما في العالم هذا القرن وتصبح محط أنظار كل العلماء والمهتمين بالبيانات ومنها أساليب التحليل العنقودي والتي نتناولها بالتفصيل لاحقاً.

1-1 : مشكلة الدراسة:-

قديما كان يتم إختبار الأساليب الاحصائية علي بيانات معملية أو مختبرية مضبوطة حتي تتم السيطرة علي النتائج ووضعها محط الثقة.

لم يسبق للعلماء ان جربوا الإختبارات الاحصائية علي بيانات بكم هائل مثل البيانات الضخمة التي ظهرت في مطلع القرن الواحد والعشرون مما أثار تساؤل حول مدي جدوى استخدام اختبارات لم تجرب علي بيانات بهذا الكم الهائل, وصار السؤال هو ماهي الطريقة الانسب للتحليل الاحصائي عندما يزيد حجم العينة الي احجام ضخمة تفوق حجم البيانات التي اجريت عليها الاختبارات في القرن السابق بملايين المرات, مما يقودنا الي ضرورة الاجابة علي التساؤلات الاتية:

- في التحليل التصنيفي هل تطبيق التحليل العنقودي باستخدام طريقة الالمتوسطات (K-means) يوفر نتائج في زمن قياسي بدلا من الطريقة الهرمية (Hierarchical) عند الزيادة المضطردة لحجم العينة؟
- وهل طريقة المتوسطات (k-means) لها القدرة علي تصنيف امثل في مجموعات حسب الصفات المشتركة فيما بينها في هذه الاحجام للبيانات باختيار عشوائي لعدد (k) مجموعة).
- ما هي الطريقة الانسب للتحليل العنقودي اذا كانت موارد الاجهزة المستخدمة في التحليل متواضعة ولماذا؟

1-2 أهمية الدراسة:-

إن دعم المعلومات بالبحوث يساعد كثيرا في حل جميع المعضلات , ونجد ان هناك دراسات كثيرة قامت باستخدام احدي طرق التحليل العنقودي في تحليل البيانات ومعرفة عدد العناقيد الحالات التي تنتمي اليها والصفات المشتركة فيما بينها ولكن هناك ندرة في الدراسات التي تقارن طرق التصنيف حسب ترشيدها للموارد المتاحة وحسب سرعتها في اعطاء النتائج , فطرق التحليل العنقودي وكل طرق استكشاف البيانات لا تختلف في النتائج كثيرا كإختلافها في مضمونها وطرق عملها لذلك تحتم علينا أن ننظر الي هذه الطرق بنظرة جديدة وهي مدي ملائمة هذه الطرق للوضع الحالي في عالمنا ولتلبية احتياجاته المتزايدة من المعلومات الناتجة من استكشاف البيانات هنا تأتي

أهمية هذه الدراسة كدراسة مقارنة لطرق التصنيف في التحليل العنقودي باعتبارها اهم الطرق ملائمة لهذا الوضع.

1-3 أهداف الدراسة:-

تقوم هذه الدراسة علي عدد من الاهداف والتي من اجلها تم اعداده وهي:

1- التعرف علي اسلوب الطريقة الهرمية باستخدام برنامج (R-Programming).

2- التعرف علي اسلوب طريقة المتوسطات باستخدام برنامج (R-Programming).

3- تحديد اي الطريقتين انسب في حالة العينات صغيرة الحجم ولماذا.

4- تحديد اي الطريقتين انسب في حالة العينات كبيرة الحجم ولماذا.

5- مقارنة الطريقتين واستخراج الفروقات في الاداء النتاجة عن زيادة حجم العينة الي احجام ضخمة.

1-4 فرضيات الدراسة:-

1. متوسط الزمن الذي تتبعه الطريقتين يتبع التوزيع الطبيعي.
2. زيادة حجم العينة تقلل من كفاءة الطريقتين.
3. طريقة المتوسطات اقل من الطريقة الهرمية في سرعة اعطاء النتائج.
4. الطريقة الهرمية تعطي نتائج افضل من طريقة المتوسطات لمراحل تكوين العناقيد.
5. عند تحليل البيانات الكبيرة فان الطريقة المتوسطة تعطي نتائج افضل.
6. الطريقة الهرمية تساعد الباحثين في توضيح.
7. لا توجد فروق ذات دلالة احصائية في وسيط الزمن اللازم لاعطاء النتائج لطريقة المتوسطات مقارنة بالهرمية.

1-5 بيانات الدراسة:-

تم استخدام بيانات مجهزة لاغراض البحث العلمي من موقع مركز تعليم الآلة والأنظمة الذكية[2]. التابع لجامعة كاليفورنيا هي جامعة بحثية عامة تقع في إرفين، كاليفورنيا، الولايات المتحدة الأمريكية، والذي يوفر حزم بيانات حقيقية مجهّزة لاغراض البحث العلمي مأخوذة من مصادرة جمع البيانات, تم جمع البيانات في الفترة بين 1999 – 2008 من عدد (130) مستشفي في الولايات المتحدة , وتقدم البيانات بالنيابة عن مركز البحوث السريرية والبحوث متعدية الجنسيات،

[2]https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008
15/11/2016 - 08:30am

التابع جامعة فرجينيا كومنولث، المستفيدة من المعاهد الوطنية للصحة ولقد تم إعداد هذه البيانات لتحليل العوامل والنتائج المتعلقة بالمرضى الذين يعانون من مرض السكري وتصنيفهم , البيانات تخص عدد (101766) مريض.

1-6 منهجية الدراسة:

في هذه الدراسة سوف يتم استخدام الأسلوب الوصفي عن طريق وصف متغيرات الدراسة واستخدام منهج التحليل العنقودي عن طريق تحليل البيانات و تصنيف المرضي حسب درة تجانسهم وتطبيق طريقتي التصنيف (الهرمية والمتوسطات) وسوف يتم استخدام عدد من البرامج هي:-

1. برنامج الحزمة الإحصائية للعلوم الإجتماعية (SPSS) (النسخة IBM SPSS 20)
2. برنامج (Programming-R) (النسخة R i386 3.3.2)
3. برنامج (MS-Excel 2010).

1-7 البحوث والدراسات السابقة:-

سوف نستعرض فيما يلي بعض البحوث والدراسات السابقة والاوراق العلمية التي تم فيها تناول موضوع الدراسة بصورة جزئية أو كلية :

1. د.فيصل ناجي نامق , "**دراسة تحليلية مقارنة للاعوام 2006 , 2007 , 2008 لتصنيف محافظات العراق وفقا لاصابة مرض الكبد الفيروسي باستخدام التحليل العنقودي**", الكلية التقنية الادارية , بغداد .

تلخصت هذه الدراسة في التحدث عن خطورة مرض الكبد الفيروسي علي صحة الانسان ولكثرة الاصابات في كافة القُطر دعت هذه الاسباب لاجراء دراسة تحليلية مقارنة للسنوات (2006 – 2007 – 2008) ومن ثم تصنيف محافظات القطر وفقا لخصائص مشتركة تتعلق بهذا المرض من خلال استخدام التحليل العنقودي الذي هدف الي اكتشاف نمط معين ينظم الافراد المصابين بالانواع الاكثر انتشارا لمرض الكبد الفيروسي ومن ثم تقسيم المحافظات الي عناقيد تتمتع عناصرها بخصائص مشتركة .

وتلخصت نتائجها في :

ان تباعد وتقارب قيم المعاملات في جدول التقارب يحدد عدد العناقيد التي تتكون تلقائيا وطريقة دمج العناقيد بطريقة (linkage, group, between) تعتمد علي المتوسط الاقل مسافة بين كافة الازواج ومن ثم يتم دمج المجموعتين الاكثر قربا وان هذه المسافات قد اختلفت من سنة الي اخري .

2. محمد بكري عبيد,(2015)، "**تحديد العوامل المؤثرة في مرض السكري بإستخدام طرائق متعددة المتغيرات**", جامعة السودان للعلوم والتكنولوجيا

هدفت الدراسة الي تحديد اهم العوامل المؤثرة علي الاصابة بمرض السكري وأهم مضاعفات مرض السكري باستخدام اساليب متعدد المتغيرات ومنها التحليل العنقودي حيث خلصت الاطروحة الي تصنيف مرض السكري في ولاية شمال كردفان بعد قياس 10 متغيرات له , و توزيع عوامل الاصابة بمرض السكري الي اربعة عناقيد , حيث شمل العنقود الاول علي (الاصابة بامراض وراثية ,الاصابة بامراض العيون , الاصابة بامراض اخري) والعنقود الثاني (امراض القلب وامراض الكلي والجهاز العصبي وجروح اليد وجروح الرجل) وضم العنقود الثالث (المضاعفات فقط) واحتوي العنقود الاخير علي (بتر الاطراف فقط).

3. فرح عبدالله محمد ,(2010) ،"**تصنيف الولايات السودانيه ذات الخصائص الديمغرافية المتشابهة بإستخدام التحليل العنقودي للعام 2002 م**" , جامعة السودان للعلوم والتكنولوجيا.

تناولت هذه الدراسة تجميع الولايات السودانية في مجموعات متجانسة داخليا من حيث المتغيرات الديمغرافية وفتح الباب أمام أبحاث متشابهة لتجميع الولايات على أساس أكبر من المتغيرات . وقاد تم إستخدام طريقتين للتجميع وهما طريقة التحليل العنقودي متوسطات . وقاد أعطت K والهرمي والتحليل العنقودي ذو الطريقتان نتائج متقاربة.

ومن النتائج الهامة التي توصلت إليها الدراسة أن هناك تماثل بدرجة كبيرة من حيث الخصائص الديمغرافية بين ولايتي سنار وغرب دارفور من جهة وبين ولايتي غرب كردفان وشمال كردفان من جهة أخرى من حيث الخصائص الديمغرافية ، كذلك فإن ولاية البحر الاحمر لم تضم اي مجموعة (أي بمعنى أنها غير متماثلة مع أي مفردة (اولية) أو مجموعة إل في المرحلة قبل الاخيرة من عملية الدمج) أما ولاية الخرطوم فلم تدمج مع أي مفردة أو مجموعة في المرحلة الأخيرة .

وتعتبر هذه الدراسة مفيده للجهات التي تسعى لتجميع الولايات السودانية المتشابهة ديمغرافيا.

4. مزمل الناير سومي ,(2015) ، "**تحليل امانيات التنمية الاقليمية في السودان باستخدام التحليل العاملي والعنقودي**", جامعة السودان للعلوم والتكنولوجيا

قام الباحث باستخدام التحليل العنقودي لتجميع ولايات السودان في اقاليم حيث خلصت الدراسة الي تجميع ولايات السودان في 4 اقاليم تنموية حيث احتوي العنقود الاول او الإقليم الأول:

ضم ولاية الخرطوم . الإقليم الثاني: ضم ولاية الجزيرة الإقليم الثالث :ضم ولاية جنوب كردفان ، جنوب دارفور ، غرب دارفور ، شمال كردفان.

الإقليم الرابع: ضم ولايات كسلا ، القضارف ، النيل الأبيض ، سنار ، النيل الأزرق ، شمال دارفور ، البحرالأحمر ، نهر النيل ، الشمالية .

5. علي عبدالحافظ إبراهيم ،(2008)، "**استخدام طريقتي تحليل مقياس متعدد الابعاد والتحليل العنقودي لتحليل مجموعة من الاواني الفخارية اكتشفت في الفترة ما قبل الميلاد**"،كلية العلوم ، مجلة جامعة النهرين

هدفت هذه الدراسة الي تصنيف وتمييز عدد من القطع الفنية والاثرية من خلال دراسة الاشكال الهندسية الخاصة بكل قطعة منها علي حده باستخدام طريقتي تحليل مقياس متعدد الابعاد والتحليل العنقودي لتصنيف (25) نموذج من نماذج الاواني الفخارية المختلفة .

كانت المخرجات كالاتي :

- المجموعة الاولي , التي احتوت علي اكبر عدد من النماذج (نسبة 60% من العدد الكلي) تميزت هذه النماذج بكونها من الحجوم الكبيرة .
- المجموعة الثانية , تالفت من ثاني اكبر عدد من النماذج (بنسبة 16 % من العدد الكلي) تميزت هذه النماذج بكونها من الحجم المتوسط وذات فتحات عليا واسعة .
- المجموعة الثالثة , احتوت اقل عدد من النماذج (12 % من العدد الكلي) تميزت هذه النماذج بامتلاكها لحجم متوسط ولكن بفتحات ضيقة مقارنة بالمجموعة الثانية
- المجموعة الرابعة , احتوت اقل عدد من النماذج (12 % من العدد الكلي) تميزت هذه النماذج بامتلاكها نماذج صغيرة الحجم .

وكانت اهم الاستنتاجات من هذه المخرجات كالاتي :

تطابق وتوافق نتائج الطريقتين المستخدمة في الدراسة , حيث صنفت الاواني الي نفس العدد من المجاميع من جهة , وقدد تماثلت في تصنيف نفس العدد من النماذج في داخل المجموعة الواحدة.

ايضا تمكنت طريقة تحليل المقياس متعدد الابعاد من تحديد تفاصيل ادق من طريقة التحليل العنقودي من خلال تحديد وتصنيف النماذج علي خارطة ذات بعدين.

التعليق علي الدراسات السابقة :

في الدراسة الاولي تعرضت الدراسة الي طريقة دمج المحافظات في عناقيد حسب درجة الاصابة بالمرض ولكن لم دراسة هل الطرق التي استخدمت في تكوين العناقيد هي الامثل او الاسرع في اعطاء النتائج ، حيث تعرض الباحث في توصياته الي ضرورة توفير الادوية والعلاجات اللازمة للمرض والاهتمام به ولم يتم التعرض لانتقاد طرق التحليل العنقودي في كيفية عملها.

تناولت الدراسة الثانية كافة الطرق لتحيل متعدد المتغيرات من التحليل العاملي والتحليل التمييزي والتحليل العنقودي كلها في دراسة واحدة حيث صار من الصعب التركيز علي اداء طريقة معينة من الطرق داخل احد تلك التصانيف مما يجعل التطرق لتفاصيل تلك الطرق ضعيف ، حيث تم التركيز علي النتائج النهائية لأي طريقة وانتقاد نواقصها دون التطرق لاسباب هذا القصور في اعطاء تفاصيل اكثر عن اسلوب التصنيف.

كما ركز الباحث في الدراسة الثالثة علي طريقة المتوسطات والطريقة الهرمية من حيث تكوين العناقيد وقد تطرق الي مزايا داخلية تجمع الولايات مع بعضها وجعلها متجانسة داخليا دون التطرق لعامل المقارنة بين الطريقتين حيث تم انتقاد عمل كل طريقة دون اعتماد معيار محدد لتحديد ايهما افضل.

وتطرق الباحث في الدراسة الرابعة الي تصنيف (15) ولاية من ولايات السودان الي عناقيد من حيث امكانيات التنمية الاقليمية ولكن تناول في توصياته انه "كلما قل حجم الوحدة الاحصائية المستخدمة في التحليل (كانت الولايات في هذه الدراسة) كلما ارتفعت درجة الدقة الاحصائية وبالتالي دقة النتائج" .

وما تواجهه هذه التوصية انه لا مجال لتطبيقها مستقبل مع حجم البيانات التي نتحدث عنها في هذه الدراسة مما يجعل افتراض قلة الوحدة الاحصائية لا ينطبق في العالم الحديث حيث اصبحت الوحدات الاحصائية تصل الي الآلاف والملايين من الحالات والمتغيرات.

نلاحظ في الدراسة الخامسة قلة وندرة هذا النوع من البيانات التاريخية عن الآنية الفخارية وبالتالي كان المجال ضيق ولا تنطبق فيه شروط البيانات الكبيرة حيث تعتبر طرق التحليل العنقودي من اهم الطرق المستخدمة في استكشاف البيانات الكبيرة غير معلومة التصنيف مسبقاً.

1-8 هيكل الدراسة :-

اشتمل البحث علي فصول حيث احتوي الفصل الاول علي المقدمة وتشمل منهجية الدراسة ويبين الفصل الثاني تفرع التحليل العنقودي ومن اين ينحدر وشمل الفصل الثالث علي تفصيل لطريقتي المتوسطات والهرمية المستخدمة في الدراسة واحتوي الفصل الرابع علي نتائج التجربة العملية علي حزمة البيانات وختاما الفصل الخامس فيه توصيات الدراسة والمقترحات التي تصاحبها ثم المراجع والملاحق.