The North African Journal of Scientific Publishing (NAJSP)
مجلة شمال إفريقيا للنشر العلمي (NAJSP)
E-ISSN: 2959-4820
Volume 4, Issue 4, 2026
Page No: 94-103
Website: https://najsp.com/index.php/home/index
SJIFactor 2024: 5.49 معامل التأثير العربي (AIF) 2025: 0.69 ISI 2024: 0.696

The Role of Statistical Methods in Large-Scale Data Analysis within Data Science: A Theoretical and Applied Study

Salem Abdulali Ahmed Abdulali[1*], Ali Mohamed Ali Ehmida[2]
[1,2]General Department, Faculty of Economics, Misrata University, Misrata, Libya

دور الأساليب الإحصائية في تحليل البيانات واسعة النطاق ضمن علم البيانات: دراسة نظرية وتطبيقية

سالم عبد العالي أحمد عبد العالي[1*]، علي محمد علي حميدة[2]
[1،2]القسم العام، كلية الاقتصاد، جامعة مصراتة، مصراتة، ليبيا

*Corresponding author: s.abdulali@eps.misuratau.edu.ly

Received: August 01, 2026 | Accepted: October 12, 2026 | Published: October 27, 2026

Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

Abstract:
This research aims to highlight the role of statistical methods in the analysis of large-scale data within the framework of modern data science. With rapid advances in technologies for collecting, storing, and processing data, big data has become one of the most prominent challenges and opportunities for researchers and practitioners. The study adopts a mixed approach that combines a theoretical review of the literature with an applied analysis of the 1994 U.S. Adult Census Income dataset, using its training set of 32,561 observations and 14 demographic and economic variables. It should be noted that this dataset does not constitute big data in the full sense of the term; rather, it is a relatively large benchmark dataset selected as an applied example to illustrate the use of statistical methods and the issues that arise with large samples. The theoretical part reviews the concept of big data and its characteristics (Volume, Velocity, Variety, Veracity, and Value), together with the main statistical methods used in the field, including descriptive, inferential, and predictive approaches. The applied part presents a statistical analysis comprising descriptive statistics, the independent-samples t-test for comparing means, the chi-square test of independence for categorical variables, and correlation analysis, supplemented by effect size measures (Cohen's d and Cramer's V). At a significance level of $\alpha = 0.05$, the results showed statistically significant differences ($p < 0.001$) in mean age and weekly working hours between the two income groups, with medium effect sizes, as well as a statistically significant association between education level and income with a medium effect size (Cramer's V = 0.37), whereas the correlations among the quantitative variables were weak. The study concludes that statistical methods are a fundamental pillar of data science, and that applying them correctly, with attention to effect size rather than statistical significance alone, improves the accuracy of analysis and the reliability of data-driven decisions.

Keywords: Large-Scale Data, Big Data, Data Science, Statistical Methods, Descriptive Statistics, Inferential Statistics, Effect Size, Predictive Analysis, Classification Models.

الملخص:
يهدف هذا البحث إلى تسليط الضوء على الدور المحوري الذي تؤديه الأساليب الإحصائية في تحليل البيانات واسعة النطاق ضمن إطار علم البيانات الحديث. ففي ظل التطور المتسارع في تقنيات جمع البيانات وتخزينها ومعالجتها، أصبحت البيانات الضخمة تمثّل أحد أبرز التحديات والفرص في آنٍ واحد أمام الباحثين والممارسين. ويعتمد البحث على منهجية مختلطة تجمع بين التحليل النظري لأدبيات المجال والتطبيق العملي على مجموعة بيانات دخل التعداد السكاني الأمريكي لعام 1994 (Adult Census Income Dataset)، إذ استُخدمت مجموعة التدريب منها المكوّنة من 32,561 مشاهدة و14 متغيراً ديموغرافياً واقتصادياً. وتجدر الإشارة إلى أن هذه البيانات لا تمثّل بيانات ضخمة. (Big Data) بالمعنى الكامل للمصطلح، وإنما هي مجموعة بيانات كبيرة نسبياً اختيرت نموذجاً تطبيقياً لتوضيح كيفية توظيف الأساليب الإحصائية والتحديات المرتبطة بالعينات الكبيرة. يتناول الجانب النظري مفهوم البيانات الضخمة وخصائصها (الحجم والسرعة والتنوع والصحة والقيمة)، ويعرض طيفاً واسعاً من الأساليب الإحصائية المستخدمة في هذا المجال من إحصاء وصفي واستنتاجي ونماذج تنبؤية. أما الجانب التطبيقي فيقدّم تحليلاً إحصائياً يشمل الإحصاء الوصفي، واختبار t للعينات المستقلة لمقارنة المتوسطات، واختبار كاي تربيع لاستقلالية المتغيرات النوعية، وتحليل الارتباط، مع حساب مقاييس حجم الأثر (Cohen's d وCramér's V). وعند مستوى دلالة α = 0.05، أظهرت النتائج وجود فروق دالة إحصائياً (p < 0.001) في متوسطي العمر وساعات العمل بين فئتي الدخل بحجم أثر متوسط، كما كشفت عن ارتباط دال إحصائياً بين المستوى التعليمي ومستوى الدخل بحجم أثر متوسط (Cramér's V = 0.37)، في حين جاءت معاملات الارتباط بين المتغيرات الكمية ضعيفة. وخلص البحث إلى أن الأساليب الإحصائية تمثّل ركيزة أساسية لعلم البيانات، وأن توظيفها الصحيح، مع الاهتمام بحجم الأثر لا بالدلالة الإحصائية وحدها، يُسهم في تحسين دقة التحليل وموثوقية القرارات المبنية على البيانات.

الكلمات المفتاحية: البيانات واسعة النطاق، البيانات الضخمة، علم البيانات، الأساليب الإحصائية، الإحصاء الوصفي، الإحصاء الاستنتاجي، حجم الأثر، التحليل التنبؤي، نماذج التصنيف.

المقدمة:

يشهد عصرنا الحالي ثورة حقيقية في مجال البيانات، إذ تُولِّد المؤسسات والأفراد والأجهزة المتصلة بالإنترنت كميات هائلة من البيانات بشكل يومي. وتشير التقديرات إلى أن البشرية تُنتج نحو 2.5 كوينتيليون بايت من البيانات يومياً، وهو رقم ينمو باطّراد مع التوسع في استخدام تقنيات إنترنت الأشياء والذكاء الاصطناعي والشبكات الاجتماعية (Dobre & Xhafa, 2014). هذا الكم الهائل من البيانات - المعروف بـ "البيانات الضخمة" - يفتح آفاقاً واسعة للاكتشاف والتحليل والتنبؤ، غير أنه يطرح في الوقت ذاته تحديات منهجية جسيمة على صعيد الاستخلاص والتحليل الموثوق.

في هذا السياق، برز علم البيانات (Data Science) بوصفه تخصصاً متعدد التخصصات يجمع بين الإحصاء والرياضيات وعلوم الحاسوب وخبرة المجال، بهدف استخلاص المعرفة والرؤى القابلة للتطبيق من البيانات الضخمة (Fawcett, 2013 & Provost). وتحتل الأساليب الإحصائية مكانة محورية في هذا التخصص، إذ توفر الأدوات الرياضية اللازمة لوصف البيانات واستكشافها، واختبار الفرضيات، وبناء النماذج التنبؤية، والاستدلال العلمي الصحيح من العينات إلى المجتمعات الكبيرة.

ومن هذا المنطلق، يسعى هذا البحث إلى تقديم رؤية شاملة ومتكاملة للدور الذي تؤديه الأساليب الإحصائية في تحليل البيانات الضخمة، من خلال دراسة نظرية معمّقة تُرافقها دراسة تطبيقية على بيانات تعداد سكاني حقيقية، بما يُسهم في ردم الهوة بين الإطار النظري والتطبيق العملي في هذا المجال.

مشكلة البحث وأهميته:

تكمن مشكلة البحث في الفجوة القائمة بين النمو الهائل في حجم البيانات الضخمة من جهة، وضعف استيعاب الباحثين والممارسين للأساليب الإحصائية الملائمة للتعامل مع هذا النوع من البيانات من جهة أخرى. فكثير من الدراسات تستخدم أساليب إحصائية تقليدية صُممت للتعامل مع بيانات صغيرة الحجم، دون مراعاة الخصائص الجوهرية للبيانات الضخمة كعدم الاستقرار والتوزيع غير الطبيعي والارتباطات المعقدة بين المتغيرات (Fan et al,.2014).

أهمية البحث:

1. **تعزيز الفهم النظري:** يُقدم البحث إطاراً نظرياً متكاملاً يربط بين مفاهيم البيانات الضخمة وأدوات علم البيانات والأساليب الإحصائية.
2. **الأهمية التطبيقية:** يُجسّد البحث كيفية تطبيق الأساليب الإحصائية على بيانات واقعية واسعة النطاق، مما يُثري الجانب التطبيقي للمجال.
3. **سد الفجوة المعرفية:** يُسهم في تقليص الهوة بين التخصصات التقنية (علوم الحاسوب) والمنهجية (الإحصاء) ضمن منظومة علم البيانات.
4. **دعم اتخاذ القرار:** تُقدم نتائجه مرجعاً عملياً للمؤسسات والهيئات الساعية إلى توظيف البيانات في تحسين قراراتها.
5. **الإسهام الأكاديمي:** يُثري الأدب العلمي العربي في مجال إحصاء البيانات الضخمة الذي يفتقر إلى الدراسات الشاملة باللغة العربية.

أهداف البحث:

1. تحليل مفهوم البيانات الضخمة وخصائصها وتحدياتها الإحصائية بصورة منهجية.
2. استعراض الأساليب الإحصائية المستخدمة في علم البيانات وتصنيفها وفق معايير موضوعية.
3. تطبيق مجموعة متنوعة من الأساليب الإحصائية على بيانات التعداد السكاني الأمريكي (Adult Dataset).
4. قياس الفروق الإحصائية بين الفئات الديموغرافية المختلفة في علاقتها بمستوى الدخل.
5. تقديم توصيات عملية لاختيار الأساليب الإحصائية المناسبة وفق طبيعة البيانات وأهداف التحليل.

أسئلة البحث:

1. ما الأساليب الإحصائية الأنسب لتحليل البيانات الضخمة ضمن إطار علم البيانات الحديث؟
2. هل توجد فروق ذات دلالة إحصائية في متوسطات العمر وساعات العمل الأسبوعية بين فئتي الدخل المرتفع والمنخفض في بيانات التعداد السكاني؟
3. ما طبيعة العلاقة بين المستوى التعليمي ومستوى الدخل، وهل هما مستقلان أم مرتبطان إحصائياً، وما حجم هذا الارتباط؟
4. كيف يُسهم الإحصاء الاستنتاجي في تعزيز موثوقية الاستنتاجات المستخلصة من العينات الكبيرة؟

منهج البحث:

يعتمد هذا البحث على المنهج الوصفي التحليلي المدعوم بالتطبيق الكمي، إذ يُجمع بين أسلوبين منهجيين متكاملين: الأسلوب الاستنتاجي في الجانب النظري، حيث يُستخلص من الأدبيات العلمية والدراسات السابقة إطار نظري متكامل، والأسلوب الكمي في الجانب التطبيقي، من خلال تطبيق حزمة متنوعة من الأساليب الإحصائية الوصفية والاستنتاجية على بيانات التعداد السكاني الأمريكي لعام 1994 (Adult Census Income Dataset) المستخرجة من قاعدة بيانات UCI للتعلم الآلي.

تتكون مجموعة البيانات المستخدمة من 32,561 سجلاً في مجموعة التدريب، يصف كلٌّ منها فرداً بواسطة 14 متغيراً تشمل المتغيرات الكمية (كالعمر، وسنوات التعليم، وساعات العمل الأسبوعية، ومكاسب رأس المال) والمتغيرات النوعية (كالجنس، والعرق، والمهنة، والحالة الاجتماعية). ويتمثل المتغير التابع في تصنيف ثنائي لمستوى الدخل (≥K50 أو <K50 دولاراً سنوياً). وقد أُجري التحليل الإحصائي باستخدام لغة Python مع مكتبات Pandas وNumPy وSciPy.

وقد اقتصر التحليل على مجموعة التدريب (adult. Data) البالغة 32,561 مشاهدة، دون مجموعة الاختبار (adult. Test) البالغة 16,281 مشاهدة؛ لأن تقسيم البيانات إلى مجموعتي تدريب واختبار صُمّم أصلاً لتقييم أداء نماذج التعلم الآلي التنبؤية على بيانات لم تُستخدم في بنائها، في حين يقتصر هدف الجانب التطبيقي في هذا البحث على التحليل الإحصائي الوصفي والاستنتاجي الذي لا يتطلب عينة اختبار مستقلة. فضلاً عن أن حجم مجموعة التدريب كافٍ لتحقيق قوة إحصائية مرتفعة جداً لجميع الاختبارات المستخدمة.

واعتُمد مستوى الدلالة $\alpha = 0.05$ في جميع الاختبارات الإحصائية، وتُعرض قيم p الصغيرة جداً بصيغة موحّدة هي $p < 0.001$. ونظراً لأن العينات الكبيرة تجعل أصغر الفروق دالة إحصائياً، حُسب حجم الأثر لتقدير الأهمية العملية للنتائج؛ إذ استُخدم معامل Cohen’s d مع اختبار t، ومعامل Cramer's V مع اختبار كاي تربيع، وفُسِّرت قيمهما وفق التصنيف الشائع لـ Cohen (1988): لمعامل d: 0.2 صغير، و0.5 متوسط، و0.8 كبير؛ ولمعامل V (في الجداول التي يكون أحد بُعديها فئتين): 0.1 صغير، و0.3 متوسط، و0.5 كبير.

هيكل البحث:

يتكون هذا البحث من مقدمة وخاتمة يتوسطهما مبحثان رئيسيان: يُعنى المبحث الأول بالإطار النظري الذي يستعرض مفهوم البيانات الضخمة وعلم البيانات والأساليب الإحصائية المستخدمة في تحليلها مع استعراض الدراسات السابقة ذات الصلة. أما المبحث الثاني فيتضمن الدراسة التطبيقية التي تُجري تحليلاً إحصائياً متكاملاً لبيانات التعداد السكاني، يشمل الإحصاء الوصفي والاستنتاجي وتحليل الارتباط واختبار الفرضيات، مع عرض النتائج ومناقشتها في ضوء الأدبيات النظرية.

المبحث الأول: الإطار النظري:

البيانات الضخمة وعلم البيانات والأساليب الإحصائية:

مفهوم البيانات الضخمة وخصائصها:

ظهر مصطلح "البيانات الضخمة" (Big Data) في الأدبيات العلمية مطلع القرن الحادي والعشرين، وإن كانت جذوره الفكرية تعود إلى عقد التسعينيات حين بدأ الباحثون يُدركون عجز الأدوات التقليدية عن معالجة البيانات المتنامية. لا يوجد تعريف موحد ودقيق للبيانات الضخمة في الأدبيات، غير أن التعريف الأوسع قبولاً هو ذلك الذي يصفها بأنها: "مجموعات البيانات الضخمة والمعقدة التي تتخطى قدرة الأدوات التقليدية لمعالجة البيانات على التعامل معها بكفاءة" (Laney, 2001; Gartner, 2012).

النموذج الخماسي لخصائص البيانات الضخمة (5Vs):

يُمثّل النموذج الخماسي الإطار المرجعي الأكثر شيوعاً لوصف خصائص البيانات الضخمة (Chen et al., 2014; Hashem et al., 2015)، ويتضمن الأبعاد الخمسة التالية:

1. **الحجم (Volume)**: يشير إلى الكميات الهائلة من البيانات المتولدة والمخزنة، والتي تتراوح بين البيتابايت والزيتابايت. يتطلب هذا الحجم بنى تحتية موزعة وخوارزميات إحصائية مُصمَّمة للتعامل مع العينات الكبيرة.
2. **السرعة (Velocity)**: تُقاس سرعة توليد البيانات والحاجة لمعالجتها في الوقت الحقيقي أو شبه الحقيقي. تفرض هذه السرعة اشتراطات خاصة على الأساليب الإحصائية من حيث الكفاءة الحسابية.
3. **التنوع (Variety)**: تأتي البيانات الضخمة في أشكال متعددة: منظمة (جداول قواعد البيانات)، وشبه منظمة (XML, JSON)، وغير منظمة (نصوص، صور، مقاطع مرئية). يتطلب كل نوع أساليب إحصائية مختلفة.
4. **الصحة (Veracity)**: تعكس مستوى الضجيج والتناقض وعدم اليقين في البيانات. تُعد إدارة جودة البيانات وتنظيفها من أبرز التحديات الإحصائية في هذا المجال (Hariri et al., 2019).
5. **القيمة (Value)**: يُمثّل هذا البُعد الهدف الغائي؛ أي القدرة على استخلاص رؤى ذات قيمة تجارية وعلمية وسياسية من الكتلة الضخمة من البيانات الخام.

علم البيانات وعلاقته بالإحصاء:

يُعرَّف علم البيانات بأنه "مجال متعدد التخصصات يوظّف الأساليب العلمية والعمليات والخوارزميات والأنظمة لاستخراج المعرفة والرؤى من البيانات المنظمة وغير المنظمة" (Donoho, 2017). يقع الإحصاء في صميم علم البيانات، إذ يُزوّده بنظرية الاستدلال، وأدوات التقدير، وأساليب التحقق من الفرضيات، وأُطر عدم اليقين الضرورية لتحويل البيانات إلى معرفة موثوقة.

يرى Cleveland (2001) في ورقته الرائدة "Data Science: An Action Plan" أن الإحصاء يجب أن يتوسع ليشمل الحوسبة ومعالجة البيانات الكبيرة، في حين يرى Breiman (2001) أن ثمة ثقافتين في التحليل الإحصائي: ثقافة النمذجة الإحصائية التقليدية، وثقافة التنبؤ الحسابي (التعلم الآلي)، مؤكداً أن كليهما ضرورة لا غنى عنها في علم البيانات.

الأساليب الإحصائية في تحليل البيانات الضخمة:

يمكن تصنيف الأساليب الإحصائية المستخدمة في تحليل البيانات الضخمة ضمن ثلاثة محاور رئيسية:

أ. **الإحصاء الوصفي (Descriptive Statistics):**

يُعنى الإحصاء الوصفي بتلخيص البيانات وعرضها بطريقة مُفيدة ومُيسَّرة الفهم. ويشمل في سياق البيانات الضخمة مجموعة من الأدوات الأساسية:

1. **مقاييس النزعة المركزية:** المتوسط الحسابي والوسيط والمنوال، مع الأخذ بعين الاعتبار تأثير القيم المتطرفة الشائعة في البيانات الضخمة.
2. **مقاييس التشتت:** الانحراف المعياري والتباين والمدى والمدى الربيعي، وهي مؤشرات حيوية لفهم توزيع البيانات.
3. **مقاييس الشكل:** معامل الالتواء (Skewness) ومعامل التفلطح (Kurtosis)، اللذان يكتسبان أهمية خاصة في البيانات الضخمة نظراً لشيوع الانحراف عن التوزيع الطبيعي.
4. **الرسوم البيانية والتصويرية:** المدرجات التكرارية والمخططات الصندوقية ومخططات التشتت، التي تُسهم في الكشف عن الأنماط والشذوذات.

ب. **الإحصاء الاستنتاجي (Inferential Statistics):**

يسعى الإحصاء الاستنتاجي إلى تعميم نتائج العينة على المجتمع الأصل مع تحديد مستوى الثقة المصاحب لهذا التعميم. ويتضمن الأدوات التالية ذات الصلة بالبيانات الضخمة:

1. **اختبارات الفرضيات (Hypothesis Testing):** تُمكّن من اتخاذ قرارات مبنية على الأدلة الإحصائية، مثل اختبار t ستيودنت لمقارنة المتوسطات، واختبار كاي تربيع لاستقلالية المتغيرات النوعية.
2. **تقدير الفترات (Interval Estimation):** توفير نطاقات ثقة بدلاً من تقديرات نقطية، مما يُعبّر بصدق عن عدم اليقين في الاستدلال.
3. **تحليل التباين (ANOVA):** يُقارن متوسطات أكثر من مجموعتين في آنٍ واحد، وهو مفيد في تحليل البيانات الضخمة ذات المتغيرات التصنيفية متعددة الفئات.
4. **اختبار كولموغوروف-سميرنوف (K-S Test):** يُقارن التوزيعات الإجمالية للعينات، ويُستخدم على نطاق واسع للتحقق من مدى تمثيل العينات لمجتمعاتها في البيئات الموزعة.

ت. **الأساليب التنبؤية والنمذجة الإحصائية:**

تُشكّل النماذج الإحصائية التنبؤية حجر الزاوية في تطبيقات علم البيانات على البيانات الضخمة (Varian, 2014)، **وتشمل الأساليب التالية:**

1. **الانحدار اللوجستي (Logistic Regression):** أسلوب إحصائي قوي للتصنيف الثنائي، يُفيد في التنبؤ بمتغيرات ثنائية كمستوى الدخل أو قبول الائتمان.

2. **الانحدار الخطي المتعدد (Regression Multiple Linear):** يُحلّل العلاقة بين متغير تابع كمي ومتغيرات مستقلة متعددة.
3. **أشجار القرار (Decision Trees):** خوارزمية تعلم آلي ذات أساس إحصائي، تُتيح التصنيف والانحدار مع إمكانية تفسير النتائج.
4. **الغابات العشوائية (Random Forests):** امتداد لأشجار القرار يستخدم تقنية التجميع (Ensemble) للحد من الإفراط في التخصيص (Overfitting).
5. **شبكات بايز (Bayesian Networks):** تُقدم نهجاً احتمالياً لتمثيل العلاقات بين المتغيرات، استناداً إلى نظرية بايز في الإحصاء البايزي.

تحديات تطبيق الأساليب الإحصائية على البيانات الضخمة:

يُواجه الإحصائيون وعلماء البيانات جملةً من التحديات عند التعامل مع البيانات الضخمة، **يمكن إجمالها فيما يلي:**

1. **مشكلة الدلالة الإحصائية:** مع الأحجام الهائلة للبيانات، تُصبح أي فروق ضئيلة ذات دلالة إحصائية مرتفعة (p<0.05)، حتى وإن كانت غير ذات أهمية عملية. يستوجب ذلك الاعتماد على مقاييس حجم التأثير (Effect Size) جنباً إلى جنب مع الدلالة الإحصائية (Lin et al., 2013).
2. **الارتباطات الكاذبة (Spurious Correlations):** كثرة المتغيرات في البيانات الضخمة تُفضي إلى اكتشاف ارتباطات بين متغيرات ليس بينها صلة سببية حقيقية، وتتفاقم هذه المشكلة مع ارتفاع عدد الأبعاد، وهو ما يرتبط بما يُعرف بـ "لعنة الأبعاد" (Dimensionality Curse of).
3. **التوزيع غير الطبيعي:** تفشل كثير من الاختبارات الإحصائية المعلمية في البيانات الضخمة بسبب انتهاك افتراض الطبيعية، مما يستلزم اللجوء إلى الأساليب اللامعلمية أو الأساليب المستندة إلى Bootstrap.
4. **تعدد الاختبارات (Multiple Testing):** تُحدث الاختبارات الإحصائية المتعددة خطر ارتفاع معدل الخطأ من النوع الأول (Positive False)، مما يستلزم تصحيحات إحصائية كتصحيح Bonferroni أو معدل الاكتشاف الكاذب (FDR).
5. **التحديات الحسابية:** تستلزم البيانات الضخمة خوارزميات إحصائية قابلة للتوسع (Scalable) تعتمد على المعالجة الموزعة (MapReduce, Spark) والحوسبة المتوازية، بدلاً من الأساليب الإحصائية التقليدية القائمة على العمليات المركزية.

الإطار النظري والدراسات السابقة:

تراكم في الأدبيات العلمية خلال العقد الأخير كمٌّ وافر من الدراسات التي تناولت الأساليب الإحصائية في علم البيانات والبيانات الضخمة من زوايا متعددة:

1. دراسة **Fan et al**. (2014): قدمت هذه الدراسة الرائدة المنشورة في مجلة "National Science Review" مناقشةً شاملة لتحديات الإحصاء في عصر البيانات الضخمة، مُشيرةً إلى ضرورة تطوير أساليب إحصائية جديدة تُراعي الطبيعة الموزعة وعالية الأبعاد للبيانات الضخمة.
2. دراسة **Jordan** & **Mitchell** (2015): أبرزت هذه الدراسة المنشورة في مجلة "Science" الدور المحوري للتعلم الآلي – بوصفه امتداداً للإحصاء التطبيقي – في معالجة البيانات الضخمة، وأكدت أن الأساليب الإحصائية الكلاسيكية تحتاج إلى إعادة صياغة لمواجهة تحديات هذا العصر.
3. **دراسة Manyika et al**. (2011): أصدر معهد ماكنزي هذا التقرير الشامل الذي رصد حجم البيانات الضخمة وإمكاناتها الاقتصادية، مؤكداً أن الأساليب الإحصائية والتحليلية تمثل العمود الفقري لاستثمار هذه البيانات.
4. **دراسة Sinanc** & **Sagiroglu (2013):** قدّمت مراجعة منهجية لمفهوم البيانات الضخمة وخصائصها وتقنياتها، مع التركيز على الأساليب التحليلية الإحصائية والحسابية اللازمة لمعالجتها.
5. **دراسة العمري والخشيبي (2022):** تُمثّل هذه الدراسة العربية إسهاماً مهماً في الأدب العلمي العربي، إذ استعرضت تطبيقات الإحصاء في علم البيانات في سياق المؤسسات العربية، مع التركيز على تحديات التحول الرقمي وأهمية الكوادر البشرية المؤهلة إحصائياً.

المبحث الثاني: الدراسة التطبيقية:

تطبيق الأساليب الإحصائية على بيانات التعداد السكاني:

1. **وصف مجموعة البيانات:**

تُعدّ مجموعة بيانات دخل التعداد السكاني الأمريكي (Adult Census Income Dataset) من المجموعات المرجعية الشائعة في أبحاث التعلم الآلي وعلم البيانات. وقد استخرجها Barry Becker من قاعدة بيانات مكتب الإحصاء الأمريكي (U.S. Census Bureau) لعام 1994، وأودعها Ronny Kohavi وBarry Becker في مستودع UCI للتعلم الآلي التابع لجامعة كاليفورنيا في إرفاين عام 1996 (Kohavi, 1996; Kohavi, 1996 & Becker)، وهي متاحة للعموم. تتضمن المجموعة 48,842 سجلاً إجمالاً موزعة على ملفين: مجموعة التدريب (32,561 سجلاً) ومجموعة الاختبار (16,281 سجلاً)، ويصف كل سجل فرداً بالغاً من خلال 14 متغيراً ديموغرافياً واقتصادياً، وقد استُخدمت في هذا البحث مجموعة التدريب وحدها للأسباب الموضحة في منهج البحث.

ولا تُصنَّف هذه المجموعة بيانات ضخمة بالمعنى الكامل وفق النموذج الخماسي؛ إذ يمكن معالجتها على حاسوب شخصي بالأدوات الإحصائية التقليدية، ولا تتسم بسرعة التدفق أو بتنوع الصيغ، وإنما هي بيانات واسعة النطاق نسبياً (Large-scale dataset) من حيث عدد المشاهدات. وقد اختيرت لأنها بيانات مرجعية مفتوحة وموثقة تتيح تطبيق الأساليب الإحصائية وإظهار بعض التحديات المرتبطة بالعينات الكبيرة، كتضخم الدلالة الإحصائية. ونظراً لأنها بيانات أمريكية تعود إلى عام 1994، فإن نتائجها لا تُعمَّم على المجتمعات العربية أو على الواقع الراهن، وإنما تُستخدم نموذجاً تطبيقياً لتوضيح الأساليب الإحصائية.

جدول رقم (1): وصف متغيرات مجموعة البيانات

م	اسم المتغير	الوصف	النوع
1	العمر (age)	عمر الفرد بالسنوات	كمي
2	فئة العمل (work class)	القطاع الوظيفي (حكومي/خاص/ذاتي...)	نوعي
3	الوزن النهائي (fnlwgt)	وزن إحصائي متعلق بالتعداد	كمي
4	التعليم (education)	أعلى شهادة حصل عليها الفرد	نوعي
5	سنوات التعليم (education_num)	سنوات الدراسة بالأرقام (1-16)	كمي
6	الحالة الاجتماعية (marital status)	الحالة الاجتماعية	نوعي
7	المهنة (occupation)	نوع المهنة	نوعي
8	العلاقة (relationship)	العلاقة الأسرية	نوعي
9	العرق (race)	العرق	نوعي
10	الجنس (sex)	الجنس (ذكر/أنثى)	نوعي
11	مكسب رأس المال (capital gain)	الأرباح من رأس المال	كمي
12	خسارة رأس المال (capital loss)	الخسائر من رأس المال	كمي
13	ساعات العمل (hours per week)	متوسط ساعات العمل الأسبوعية	كمي
14	بلد المنشأ (native country)	بلد المنشأ	نوعي
15	الدخل (income) - المتغير التابع	فئتان: ≥K50 أو <K50 دولاراً	ثنائي

2. **الإحصاء الوصفي:**

يُقدم الجدول التالي الإحصاءات الوصفية للمتغيرات الكمية الرئيسية في مجموعة البيانات. تجدر الإشارة إلى أن المجموعة تتضمن 32,561 سجلاً في مجموعة التدريب، منها 24,720 سجلاً (75.9%) تندرج في فئة الدخل المنخفض (≥K50)، و7,841 سجلاً (24.1%) في فئة الدخل المرتفع (<K50).

جدول رقم (2): الإحصاءات الوصفية للمتغيرات الكمية

المتغير	المتوسط الحسابي	الانحراف المعياري	الحد الأدنى	الحد الأقصى
العمر (سنة)	38.58	13.64	17	90
سنوات التعليم	10.08	2.57	1	16
ساعات العمل الأسبوعية	40.44	12.35	1	99
مكسب رأس المال	1,077.65	7,385.29	0	99,999
خسارة رأس المال	87.30	402.96	0	4,356

تُشير الإحصاءات الوصفية إلى أن متوسط عمر أفراد العينة يبلغ 38.58 سنة بانحراف معياري 13.64، كما أن متوسط ساعات العمل الأسبوعية 40.44 ساعة يتوافق مع معيار أسبوع العمل الرسمي في الولايات المتحدة. ويلفت الانتباه الانحراف الشديد في توزيع متغير مكسب رأس المال (capital gain) نحو اليمين، إذ إن غالبية الأفراد لا يمتلكون أي مكاسب رأسمالية (صفر)، في حين يمتلك أقلية منهم مكاسب ضخمة، مما يُولّد توزيعاً شديد الالتواء.

جدول رقم (3): توزيع الدخل حسب الجنس

الجنس	العدد	نسبة دخل > K50
ذكور	21,790	%30.6
إناث	10,771	%10.9
الإجمالي	32,561	%24.1

جدول رقم (4): نسبة الدخل المرتفع حسب المستوى التعليمي

المستوى التعليمي	نسبة دخل > K50
دكتوراه (Doctorate)	%74.09
دراسات مهنية (Prof-school)	%73.44
ماجستير (Masters)	%55.66
بكالوريوس (Bachelors)	%41.48
دبلوم مهني (Assoc-voc)	%26.12
بعض الكلية (Some-college)	%19.02
ثانوية عامة (HS-grad)	%15.95
ما قبل المدرسة (Preschool)	%0.00

يكشف الجدول (4) عن تدرّج واضح في نسبة الدخل المرتفع بحسب المستوى التعليمي؛ إذ تتجاوز هذه النسبة 73% لدى حاملي الدكتوراه والشهادات المهنية، في حين لا تتجاوز 8% لدى من لم يُكملوا المرحلة الثانوية. وتشير هذه النتائج الوصفية إلى وجود نمط متزايد في نسبة الدخل المرتفع مع ارتفاع المستوى التعليمي، غير أنها لا تكفي وحدها لإثبات قوة العلاقة، وهو ما يُترك للاختبار الإحصائي ومقياس حجم الأثر في القسم (2.3.2).

3. الإحصاء الاستنتاجي واختبار الفرضيات:

أ. اختبار t ستيودنت لمقارنة المتوسطات:

لاختبار وجود فروق ذات دلالة إحصائية في متوسطي العمر وساعات العمل الأسبوعية بين فئتي الدخل، يُطبَّق اختبار t ستيودنت للعينات المستقلة، مع الأخذ بعين الاعتبار اشتراط تجانس التباين واستقلالية المشاهدات، ويُقاس حجم الأثر بمعامل Cohen's d.

جدول رقم (5): نتائج اختبار t لمقارنة متوسطات العمر وساعات العمل

المتغير	متوسط ≥K50	متوسط <K50	قيمة t	قيمة p	حجم الأثر (d)	القرار الإحصائي
العمر	36.78 سنة	44.25 سنة	-43.44	< 0.001	0.56 (متوسط)	دال إحصائياً
ساعات العمل	38.84 ساعة	45.47 ساعة	-42.58	< 0.001	0.55 (متوسط)	دال إحصائياً

تكشف نتائج اختبار t عن فروق دالة إحصائياً ($p < 0.001$) في كلا المتغيرين عند مستوى دلالة 0.05. فعلى صعيد العمر، يبلغ متوسط عمر أصحاب الدخل المرتفع 44.25 سنة مقارنةً بـ 36.78 سنة لأصحاب الدخل المنخفض، أي بفارق يقارب 7.5 سنوات، وبحجم أثر متوسط ($d = 0.56$)؛ أي أن العمر الأعلى كان مرتبطاً إحصائياً بفئة الدخل المرتفع، دون أن يعني ذلك بالضرورة أن العمر يعكس سنوات الخبرة المهنية، إذ لا تتضمن البيانات متغيراً يقيس الخبرة مباشرة. أما على صعيد ساعات العمل، فيعمل أصحاب الدخل المرتفع في المتوسط 45.47 ساعة أسبوعياً مقارنةً بـ 38.84 ساعة لأصحاب الدخل المنخفض، وبحجم أثر متوسط ($d = 0.55$)، مما يشير إلى ارتباط عدد ساعات العمل بفئة الدخل، مع ضرورة الحذر من التفسيرات السببية؛ لأن الاختبار يثبت وجود فرق بين المجموعتين ولا يثبت علاقة سببية.

ب. اختبار كاي تربيع للاستقلالية:

يُطبَّق اختبار كاي تربيع (χ^2) لفحص استقلالية المتغيرات النوعية عن المتغير التابع (مستوى الدخل)، ويُقاس حجم الأثر بمعامل Cramér's V. وصيغة الفرضية الصفرية: "لا توجد علاقة بين الجنس ومستوى الدخل"، في مقابل الفرضية البديلة: "توجد علاقة ارتباطية بين الجنس ومستوى الدخل"، وتُصاغ الفرضيات بالطريقة نفسها لمتغيري التعليم والعرق.

جدول رقم (6): نتائج اختبار كاي تربيع

العلاقة المُختبَرة	قيمة χ^2	قيمة p	درجات الحرية	حجم الأثر (V)	القرار
الجنس × الدخل	1517.81	< 0.001	1	0.22 (صغير–متوسط)	رفض H_0 – علاقة دالة إحصائياً
التعليم × الدخل	4429.65	< 0.001	15	0.37 (متوسط)	رفض H_0 – علاقة دالة إحصائياً
العرق × الدخل	330.92	< 0.001	4	0.10 (صغير)	رفض H_0 – علاقة دالة إحصائياً

تُظهر نتائج اختبار كاي تربيع أن جميع العلاقات المُختبَرة دالة إحصائياً ($p < 0.001$)، ومن ثمّ تُرفض الفرضية الصفرية القائلة بالاستقلالية. غير أن اختبار كاي تربيع يثبت وجود ارتباط أو اعتماد إحصائي بين المتغيرات فحسب، ولا يدل على قوة العلاقة، لا سيما مع حجم عينة كبير يجعل أصغر الفروق دالة إحصائياً؛ لذا حُسب معامل Cramér's V لقياس حجم الأثر، فجاءت العلاقة بين التعليم والدخل متوسطة ($V = 0.37$)، وبين الجنس والدخل صغيرة إلى متوسطة

(V = 0.22)، وبين العرق والدخل صغيرة (V = 0.10). وتتسق هذه النتائج مع ما يقدّمه الأدب الاقتصادي من تفسيرات لفجوات الدخل المرتبطة بالجنس والمستوى التعليمي والعرق، غير أن التأويل السليم يستلزم التمييز بين الارتباط الإحصائي والسببية، ومراعاة العوامل المُربِكة المحتملة.

تحليل الارتباط بين المتغيرات الكمية:

يُقدّم الجدول التالي مصفوفة الارتباط (معامل بيرسون) بين المتغيرات الكمية الرئيسية. ويُفسَّر معامل الارتباط وفق التصنيف الشائع (Cohen, 1988): القيم الأقل من 0.10 تدل على ارتباط ضئيل، ومن 0.10 إلى 0.29 على ارتباط ضعيف، ومن 0.30 إلى 0.49 على ارتباط متوسط، و0.50 فأكثر على ارتباط قوي.

جدول رقم (7): مصفوفة الارتباط بين المتغيرات الكمية

المتغير	العمر	سنوات التعليم	ساعات العمل	مكسب رأس المال	خسارة رأس المال
العمر	1.000	0.037	0.069	0.078	0.058
سنوات التعليم	0.037	1.000	0.148	0.123	0.080
ساعات العمل	0.069	0.148	1.000	0.078	0.054
مكسب رأس المال	0.078	0.123	0.078	1.000	0.032-
خسارة رأس المال	0.058	0.080	0.054	0.032-	1.000

تُشير مصفوفة الارتباط إلى ارتباطات ضعيفة أو ضئيلة بين جميع المتغيرات الكمية، إذ لم يتجاوز أي معامل ارتباط القيمة 0.15، وهو أمر متوقع نظراً لتعدد العوامل المؤثرة في كل متغير. وأعلى معامل ارتباط ظهر بين سنوات التعليم وساعات العمل (r = 0.148)، وهو ارتباط ضعيف جداً لا يمكن وصفه بأنه علاقة قوية، إذ تفسّر سنوات التعليم نحو 2.2% فقط من التباين في ساعات العمل ($r^2 \approx 0.022$). وكذلك فإن الارتباط بين سنوات التعليم ومكسب رأس المال (r = 0.123) ارتباط ضعيف أيضاً ($r^2 \approx 0.015$)، ومن ثمّ لا يمكن الاعتماد عليه وحده دليلاً على نظرية رأس المال البشري، وإن كان اتجاهه يتسق معها.

مناقشة النتائج:

تُوضّح النتائج التي أسفرت عنها الدراسة التطبيقية قدرة الأساليب الإحصائية على استكشاف الأنماط داخل البيانات واسعة النطاق. فعلى صعيد التحليل الوصفي، كشف توزيع الدخل عن اختلال واضح في الفئات (75.9% دخل منخفض مقابل 24.1% دخل مرتفع)، وهو ما يُشكّل تحدياً منهجياً مهماً عند بناء النماذج التنبؤية، ويستلزم تطبيق أساليب معالجة خاصة كإعادة موازنة العينات (Oversampling/Undersampling).

وعلى صعيد الإحصاء الاستنتاجي، أظهر اختبار t وجود فروق دالة إحصائياً بحجم أثر متوسط في كلٍّ من العمر وساعات العمل بين فئتي الدخل، وهي نتيجة تتسق مع نظرية رأس المال البشري، وإن كان العمر لا يُعدّ مقياساً مباشراً للخبرة المهنية. كما أكد اختبار كاي تربيع أن الجنس والمستوى التعليمي والعرق ليست مستقلة عن مستوى الدخل، وإن تفاوتت قوة هذه العلاقات بين المتوسطة (التعليم) والصغيرة (العرق)، مما يعكس الواقع الاجتماعي-الاقتصادي لمجتمع الدراسة في تلك الفترة.

وتجدر الإشارة إلى قيود الدراسة؛ فأولاً: نظراً للحجم الكبير للعينة (32,561 مشاهدة) تصبح أصغر الفروق دالة إحصائياً، ولذلك اعتمدت الدراسة على مقاييس حجم الأثر إلى جانب الدلالة الإحصائية لتقييم الأهمية العملية للنتائج. ثانياً: البيانات أمريكية تعود إلى تعداد عام 1994، ومن ثمّ لا يمكن تعميم نتائجها على المجتمعات العربية أو على الواقع الراهن، وإنما تُستخدم نموذجاً تطبيقياً لتوضيح الأساليب الإحصائية. ثالثاً: لا تمثّل هذه البيانات بيانات ضخمة بالمعنى الكامل، إذ لا تتوافر فيها خاصيتا السرعة والتنوع، مما يستدعي تطبيق الأساليب ذاتها على بيانات ضخمة فعلية في دراسات لاحقة. رابعاً: اقتصر التحليل على مجموعة التدريب دون مجموعة الاختبار، وهو ما يتناسب مع أهداف التحليل الإحصائي الوصفي والاستنتاجي.

الخاتمة والتوصيات:

خلص هذا البحث إلى أن الأساليب الإحصائية تمثل الركيزة الجوهرية لعلم البيانات وتحليل البيانات الضخمة. فمن خلال الجانب النظري، تبيّن أن البيانات الضخمة بخصائصها الخمسة (الحجم والسرعة والتنوع والصحة والقيمة) تفرض تحديات منهجية خاصة تستلزم تطوير أساليب إحصائية مُكيَّفة تُراعي الطبيعة الموزعة وعالية الأبعاد لهذه البيانات. أما الجانب التطبيقي على بيانات التعداد السكاني الأمريكي، فقد أوضح قدرة الأساليب الإحصائية على كشف الأنماط الاجتماعية والاقتصادية في البيانات، مع أهمية تفسير النتائج في ضوء حجم الأثر لا الدلالة الإحصائية وحدها.

أبرز النتائج:

1. أسهمت الأساليب الإحصائية الوصفية في تلخيص البيانات وكشف أنماطها الأساسية، ومنها النمط المتزايد في نسبة الدخل المرتفع مع ارتفاع المستوى التعليمي.
2. كشف اختبار t للعينات المستقلة عن فروق دالة إحصائياً ($p < 0.001$) في متوسطي العمر وساعات العمل بين فئتي الدخل، بحجم أثر متوسط ($d \approx 0.55$).

3. أظهر اختبار كاي تربيع وجود علاقات دالة إحصائياً بين كلٍّ من الجنس والمستوى التعليمي والعرق ومستوى الدخل، وتراوح حجم الأثر بين الصغير (العرق) والمتوسط (التعليم).
4. أظهر تحليل الارتباط أن العلاقات الخطية بين المتغيرات الكمية جميعها ضعيفة، وأن أعلاها بين سنوات التعليم وساعات العمل (r = 0.148)، وهو ارتباط ضعيف جداً.
5. جاءت النتائج في مجملها متسقة مع نظرية رأس المال البشري القائلة بأن التعليم والخبرة من محددات الدخل، مع ضرورة الحذر في التعميم وفي التفسير السببي.

التوصيات:

1. ضرورة إدراج الأساليب الإحصائية المتخصصة في البيانات الضخمة ضمن المناهج الأكاديمية لتخصصات علم البيانات والإحصاء.
2. توظيف مقاييس حجم التأثير (Effect Size) جنباً إلى جنب مع الدلالة الإحصائية عند العمل مع عينات البيانات الضخمة.
3. الاستثمار في تطوير أدوات إحصائية قابلة للتوسع (Scalable) تستطيع معالجة البيانات الموزعة بكفاءة.
4. تشجيع الباحثين العرب على إجراء دراسات تطبيقية باستخدام البيانات الضخمة المحلية لإثراء الأدب العلمي العربي في هذا المجال.
5. التوجه نحو الأساليب الإحصائية البايزية واللامعلمية التي تُقدم مرونةً أكبر عند التعامل مع البيانات غير الطبيعية التوزيع.

قائمة المراجع:

أولاً: المراجع العربية:

1. العمري، أحمد محمد، والخشيبي، سعد بن عبد الله. (2022). الإحصاء وعلم البيانات في المؤسسات العربية: التحديات والفرص. مجلة العلوم الاقتصادية والإدارية، 28(3)، 45-72.

ثانياً: المراجع الأجنبية:

1. Becker, B., & Kohavi, R. (1996). Adult [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5XW20
2. Breiman, L. (2001). Statistical modeling: The two cultures. Statistical Science, 16(3), 199-231. https://doi.org/10.1214/ss/1009213726
3. Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. Mobile Networks and Applications, 19(2), 171-209. https://doi.org/10.1007/s11036-013-0489-0
4. Cleveland, W. S. (2001). Data science: An action plan for expanding the technical areas of the field of statistics. International Statistical Review, 69(1), 21-26.
5. Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Lawrence Erlbaum Associates.
6. Dobre, C., & Xhafa, F. (2014). Intelligent services for big data science. Future Generation Computer Systems, 37, 267-281. https://doi.org/10.1016/j.future.2013.07.014
7. Donoho, D. (2017). 50 years of data science. Journal of Computational and Graphical Statistics, 26(4), 745-766. https://doi.org/10.1080/10618600.2017.1384734
8. Fan, J., Han, F., & Liu, H. (2014). Challenges of big data analysis. National Science Review, 1(2), 293-314. https://doi.org/10.1093/nsr/nwt032
9. Gartner. (2012). Big data. Gartner IT Glossary. https://www.gartner.com/en/information-technology/glossary/big-data
10. Hariri, R. H., Fredericks, E. M., & Bowers, K. M. (2019). Uncertainty in big data analytics: Survey, opportunities, and challenges. Journal of Big Data, 6(1), Article 44. https://doi.org/10.1186/s40537-019-0206-3
11. Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Khan, S. U. (2015). The rise of "big data" on cloud computing: Review and open research issues. Information Systems, 47, 98-115. https://doi.org/10.1016/j.is.2014.07.006
12. Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255-260. https://doi.org/10.1126/science.aaa8415
13. Kohavi, R. (1996). Scaling up the accuracy of Naive-Bayes classifiers: A decision-tree hybrid. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), 202-207.
14. Laney, D. (2001). 3D data management: Controlling data volume, velocity, and variety. META Group Research Note, 6(70), 1.
15. Lin, M., Lucas, H. C., & Shmueli, G. (2013). Research commentary—Too big to fail: large samples and the p-value problem. Information Systems Research, 24(4), 906-917.
16. Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute.

17. Provost, F., & Fawcett, T. (2013). Data science for business: What you need to know about data mining and data-analytic thinking. O'Reilly Media.
18. Sagiroglu, S., & Sinanc, D. (2013). Big data: A review. 2013 International Conference on Collaboration Technologies and Systems (CTS), 42-47. IEEE.
19. Varian, H. R. (2014). Big data: new tricks for econometrics. Journal of Economic Perspectives, 28(2), 3-28. https://doi.org/10.1257/jep.28.2.3