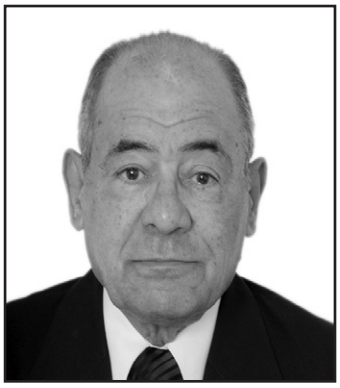
ثورة البيانات
وتحليلاتها التخطيطية والتنموية

المستخلص

يعتقد كثير من الباحثين أن البيانات الكبيرة Big data سوف تحول منظمات الأعمال والمصالح الحكومية والأبعاد الأكاديمية والعلمية وأوجه التخطيط والتنمية الأخرى للمنظمات والدول على حد سواء. ونناقش في هذا العمل مدي تأثير البيانات الكبيرة علي رسم السياسات واتخاذ القرارات التخطيطية والتنموية على كل أبعاد القطاع الخاص والقطاع العام والقطاع الحكومي. وتعتبر مجموعات البيانات الإدارية واسعة النطاق، كما ان بيانات شركات القطاع الخاص المملوكة لها يمكن من خلال تحليلها السليم والتنبئي أن تحسن الطريقة التي نقيس بها النشطة المختلفة للمنظمات والدول على حد سواء وتتبع ذلك وتضييقه بطريقة كبيرة. كما أن ظاهرة البيانات الكبيرة تساعد أيضا في تصميمات البحوث الحديثة التي تسمح للباحثين في تتبع تداعيات الأحداث أو السياسات المختلفة. ويشتمل هذا العمل علي استعراض موضوعات حاكمة تتمثل في البيانات الكبيرة فيما يتعلق بتوافرها في الوقت الحقيقي، وتوافرها علي نطاق أوسع، وعلي أنواع متغيراتها الجديدة وورودها بهيكلية أقل مما كان عليه الحال من قبل؛ بعدئذ توضح أبعاد البيانات الكبيرة وارتباطها بالنمذجة التنبئية؛ وتوضيح أساليب الإحصاء وتعلم الآلة التي تتطرق لتفسير مصطلحات علم البيانات، والاحصاء وتعلم الآلة المرتبط بالذكاء الاصطناعي وتحديد أنواع ألجوريثمات تعلم الآلة ثم التطرق للتعميم والتقييم واختيار النماذج؛وتوضيح فرص السياسات التخطيطية والاقتصادية بصفة عامة فيما يتصل باستخدام البيانات الحكومية الإدارية، وتحديد مقاييس نشاط القطاع الخاص الاقتصادي الجديدة، وتحسين عمليات وخدمات الحكومة، ومعلومات المنتجات والخدمات؛ كما إن هذا العمل يتعرض لتوضيح أبعاد الفرص المتاحة للبحوث التخطيطية والتنموية المطلوب القيام بها بالواقع المصري حيث تم مناقشة موضوعات القياس الجديد وتصاميم البحوث، وتوضيح أساليب التعلم الاحصائي والبحوث التخطيطية والتنموية وخاصة فيما يتصل احتضان عدم التجانس والاختلاف. كما تضمن هذا العمل تحديد إطار بعض التحديات التي تختص بالوصول للبيانات الكبيرة واستخدامها إلى جانب مراعاة ما لإن كانت أدوات النمذجة التنبئية للبيانات الكبيرة قد بزغت في مجالات الإحصاء وعلم الحاسب وصارت تبرهن على فائدتها القصوى في تحليلات لتخطيط والتنمية الاقتصادية المنشودة للوطن.

الكلمات الرئيسية: البيانات الكبيرة،ثورة البيانات، علم البيانات، التحليل التخطيطي، التحليل الاقتصادي، النمذجة التنبئية، التحليلات التنبئية، البيانات الإدارية، الإحصاء، تعلم الآلة.



أ.د. محمد محمد الهادي

أستاذ الحاسب الآلي ونظم المعلومات

بأكاديمية السادات للعلوم الإدارية

١. المقدمة:

الآداب المنشورة والمتاحة حاليا مكتظة بالتقارير حول كيف أن «البيانات الكبيرة Big Data» سوف تحول مجالات اعمال الشركات والمصلح الحكومية وأوجه التنمية الاجتماعية والأخرى. وقد بزغ للوجود مصطلحات حديثة مثل «علم البيانات «Data Science » و»علماء البيانات كان من الصعب السماع عنهما في السنوات القليلة الماضية. إلا أنه في عام ٢٠١٢ جادل معدوا التقارير في أن علماء البيانات قد قدموا لحملة رئيس الولايات المتحدة باراك أوباما الحد التنافسي للانتخابات الرئيسية. (Google Search for "Obama campaign" AND data mining")

كما أن الخبراء والمتخصصين الذين يعملون في وادي السيليكون Silicon valley بولاية كاليفورنيا في الولايات المتحدة الأمريكية تعجبوا بعض الوقت فيما يتصل بالتطورات الحادثة المرتبطة بثورة البيانات من خلال ظهور ظاهرة البيانات الكبيرة ومدي تأثيرها على

مخرجاتهم التكنولوجية التي سوف يكون لها مردود إيجابي على التخطيط والتنمية بأبعادها الاجتماعية والاقتصادية للشركات والدول على حد سواء.

ومن هذا المنطلق يمثل هذا العمل محاولة جادة لتقديم بعض الأفكار عن هذا التطور الذي بدأ في الظهور المتمثل في ثورة البيانات الحديثة وما أفرزته من بيانات كبيرة وتحليلات تنبئية ترتبط بها. لذلك فإن هذا العمل يتضمن بجانب هذه المقدمة عدة أجزاء رئيسية تبدأ بتحديد مفهوم وأبعاد البيانات الكبيرة من منظور المتعاملين والمتأثرين بها من مخططين واقتصاديين، وتحقيقا لذلك تم استعراض ما يحظى به مستخدمو البيانات الكبيرة من انتباه أكبر بتعلق بتعريف أنماط السوك أو النشاط وتطوير النماذج التنبئية Predictive Modeling التي تعتبر صعبة أو مستحيلة التطبيق مع عينات من مجموعات البيانات أصغر أو متغيرات أقل من البيانات التي تتسم بأنها كبيرة في الحجم والتنوع، حيث أن تنوعات تحليلات البيانات النابعة من هذه النماذج التنبئية لها تأثير رئيسي علي كثير من الصناعات التي تشتمل أيضا علي قطاعات تجارة التجزئة، التمويل، الإعلان، والتأمين الخ التي تقوم بها شركات القطاع الخاص في العادة.

كما يناقش هذا العمل كيف أن البيانات الكبيرة الجديدة قد تؤثر على مجمل أبعاد التخطيط والتنمية. ومن منظور السياسات التنموية، يلقي الضوء على قيمة مجموعات البيانات الإدارية الكبيرة، وعلى القدرة في استيعاب والتقاط بيانات العمليات المختلفة في الوقت الحقيقي وأهمية كفاءة تلك العمليات سواء النابعة من المعاملات الحكومية أو من شركات القطاع الخاص والقطاع العام. ومن منظور البحوث التنموية التي ترتبط بالتنمية الاجتماعية والاقتصادية يركز هذا العمل على كيف أن مجموعات البيانات الكبيرة يمكن أن تساعد في توضيح معالم تصاميم البحوث الجديدة ببعض الأمثلة من الأعمال الجديدة، إلى جانب تقديم امثلة يمكن للباحثين من خلالها ملاحظة تداعيات الأحداث والسياسات على مسارات التخطيط التنمية المختلفة. وفي نفس الوقت يحدد هذا العمل ما إن كانت أدوات البيانات الكبيرة قد طورت في مجالات الإحصاء وعلم الحاسب الآلي كالتعلم الاحصائي وأساليب تنقيب البيانات بهدف إيجاد التطبيقات الملائمة للتخطيط والتنمية حيث وجد أنها غير مستخدمة بكفاءة وفعالية في كثير من الأعمال سواء المتعلقة بمنظمات وشركات القطاع الخاص وفي المصالح الحكومية سواء على المستوي المركزي أو المحلي، وقد استعرض مجال الإحصاء وتعلم الالة المرتبطة بالذكاء الاصطناعي من حيث تحديد المفاهيم والأنواع والألجوريثمات والنماذج.

وفي نفس الوقت يناقش هذا العمل بعض التحديات الجديدة التي ترتبط بالبيانات الكبيرة التي تتضمن إمكانية لوصول للحصول على البيانات الحكومية في القطاع الخاص، إلى جانب الموارد الكمبيوترية الضرورية المتاحة في المجتمعات المتقدمة. وبالإضافة لذلك تظهر أهمية تدريب المخططين والقائمين بأبعاد التنمية الاقتصادية والاجتماعية للعمل مع مجموعات البيانان الكبيرة وأدوات التحليلات التنبئية المتطلبة في ذلك.

ومن الملاحظ أن هذا العمل ذا طابع تأملي ومحدود لحد كبير، وعلي وجه خاص فعلي الرغم من وصف بعض استخدامات البيانات الكبيرة الجوهرية، فلم تناقش المساوئ الأساسية التي تتعلق بالأخطار على خصوصية الأفراد أو الإقدام علي ارتكاب بعض الأشياء المحظورة التي قد تطرأ نتيجة لإمكانية استخدام البيانات الكبيرة المفصلة عن المواطنين في طرق غير مرغوب فيها. هذه القضايا المهمة التي تتصل بإنشاء وإدارة مجموعات بيانات كبيرة عن الأفراد إلا أنها لا تدخل في نطاق هذا العمل.

٢. ما البيانات الكبيرة؟

من عشرين أو ثلاثين عاما مضت، كانت البيانات عن الأنشطة الاجتماعية والاقتصادية نادرة نسبيا. وقد تغير الوضع في فترة قصيرة بصوره دراماتيكية. واحد أسباب ذلك يرجع للنمو المتنامي لشبكة الإنترنت العالمية، حيث أن كل شيء على شبكة الإنترنت مسجل عليها. وعند البحث في محركات البحث مثل Google أو Bing وغيرهما عن التساؤلات المختلفة التي يتم تسجيلها فإن ذلك يصبح مسجلا لاحقا. كما أنه عند شراء الكتب من موقع Amazon أو الأشياء الأخرى كالإلكترونيات على سبيل المثال من موقع eBay لا الشراء الذي يتم فقط بتسجل أيضا كل نقرة أي ضغطةعلى وحدات الموقع حيث تلتقط وتدون، كما انه عند قراءة إحدى الجرائد الإلكترونية أو رؤية فيديو معين أو تتبع الأحوال المالية الشخصية المتاحة علىالخط فإن كل تلك التصرفات الشخصية المعينة تسجل من قبل الموقع المحدد. كما أن تسجيل السلوك الشخصي لا يتوقف لمواقع الإنترنت فقط، بل أيضا يمكن ان يتم ايضا عند التراسل النصي من خلال الهواتف المحمولة الخلوية، وبيانات المسح، وسجلات التوظيف الإلكترونية، والسجلات الصحية الإلكترونية، والتعاملات ببطاقات الائتمان المختلفة، وبيانات أرقام الهوية أو أرقام الانتخابات الشخصية وأرقام الحسابات البنكية أو التأمينات الاجتماعية وغيرهاالتي تمثل جزءا من البيانات التي نتركها خلفنا حاليا.

وفي هذا الإطـار يمكن اعتبار مثال معين يوضح هذا الوضع الحالي للبيانات الكبيرة التي صارت متاحة في كثير من الأحيان نتيجة الثورة الرقمية التي يمر بها العالم الحديث كالبيانات المجمعة من محلات تجارة التجزئة

مثلا. ومنذ عقود قليلة مضت، كانت محلات تجارة التجزئة تحصل على بياناتها المجمعة من المبيعات اليومية ويكون ذلك بجودة عالية وخاصة عند تجزئتها بواسطة المنتجات او السلع المباعة أو مجموعاتها. أما في الوقت الحالي، فإن بيانات الماسحات Scanners صارت تجعل في الإمكان تتبع المشتريات الفردية ومبيعات السلعة او الوحدة المعينة، كما تلتقط في الوقت الحقيقي الذي تحدث فيه واقعة الشراء أو البيع وبذلك تحدد تواريخ مشتريات الأفراد، كما تستخدم بيانات المخزون الرقمية لوصل المشتريات مع مواقع ذاتية معينة لتحديد مستويات المخزون السلعي الحالية. ومن الملاحظ أن تجار التجزئة على الإنترنت لا يلاحظون هذه المعلومات فحسب، ولكنهم يمكنهم أيضا تتبع سلوك العميل حول البيع ويتضمن ذلك تساؤل البحث التمهيدي عما يقوم به كل عميل عن الوحدات التي اطلع عليها واختار شرائها أو استبعدها مما يسمح بتحديد توصيات الترويج للسلع التي تحظي بقبول أكثر، مع مراجعة السلع أو الوحدات المتاحة في المخزون. ويمكن أن ترتبط البيانات المستخلصة أساسا بالأبعاد الديموغرافية للعملاء وأنشطة الوسائل الاجتماعية، وتواريخ الائتمان، والإنفاق خارج الخط، وغير ذلك.

وفي نفس الوقت، صار يتواجد أيضا تطور متوازي يتمثل في نشاط شركات الأعمال المختلفة التي حركت عملياتها اليومية إلي الاستعانة شبه الكاملة بالحاسبات الآلية والتعامل على الخط من خلال الشبكات المتاحة لها، مما جعل ممكنا جمع بيانات غنية لتعاقدات المبيعات، وممارسات الإيجارات، ومشحونات السلع الطبيعية، الخ. وبصفة متزايدة يوجد أيضا كم كبير من السجلات الإلكترونية التي ترتبط بالعمل التعاوني، وبيئات الأفراد، وقياسات الإنتاجية. نفس الوضع يمكن ملاحظته أيضا عن مؤسسات ومنظمات القطاع العام فيما يتعلق بقدرتها على الوصول وتحليل الضرائب، وبرامج التامين الاجتماعية، والحسابات البنكية، وأبعاد القروض والمدخرات البنكية، والمصروفات الحكومية والأنشطة التشريعية وغير ذلك.

يتضح مما تقدم تواجد كميات هائلة من البيانات لدي المنظمات والشركات والمصالح الحكومية العديدة والمتشعبة في كل القطاعات. لذلك يمكننا التساؤل ما هو الجديد حول كل ذلك؟ تتمثل إجابة هذا التساؤل في ان البيانات صارت متوافرة حاليا أكثر مما كانت عليه في الماضي، كما صار لها مجالا وتغطية أعظم، وتشتمل البيانات علي أنواع كثيرة من الملاحظات والقياسات الجديدة التي لم تكن متاحة في الماضي، إلي جانب أن مجموعات البيانات الحديثة تشتمل علي هيكلية قليلة جدا وفي بعض الأحيان تكون الهيكلية معقدة اكثر، وكل ذلك يرتبط بسلاسل الوقت أو نماذج بيانات الجداول التقليدية التي صارت متاحة علي نطاق واسع أمام الباحثين في كثير من المجالات مما يسهم في تحديد أبعاد خواص البيانات الكبيرة في المعالم التالية:

١/٢ توافر البيانات في الوقت الحقيقي:

تعتبر القدرة علي التقاط ومعالجة البيانات في الوقت الحقيقي أساسية لتطبيقات كثير من الأعمال، إلا ان ذلك يستخدم لحد قليل في البحوث والسياسات المختلفة وربما يكون ذلك غير مستغربا تقريبا، حيث ان تساؤلات كثير من البحوث القائمة تعتبر تساؤلات لاستعادة الأوضاع الماضية ولا تطرق لما سوف يكون عليه الوضع مستقبلا. وبذلك يصبح من المهم أكثر للبيانات أن تفصل وتصبح دقيقة بدلا من توافرها فوريا. وسوف يستعرض في الجزء التالي من هذا العمل بعض الطرق التي تبرهن فيها بيانات في الوقت الحقيقي التي تفيد البحوث المرتبطة بالسياسات وتخطيط أبعاد التنمية.

٢/٢ توافر البيانات على نطاق واسع:

من الملاحظ أيضا أن هناك تغير رئيسي يرتبط بالقيام بالبحوث المختلفة التي يعتمد باحثوها حاليا على عينات صغيرة من مجموعات البيانات قد تتضمن مئات أو حتى آلاف المعلومات أمام الباحثين قضية مهمة. إلا انه في الوقت الحالي صارت البيانات المتاحة للباحثين تقدر بعشرات الملايين من الملاحظات المميزة التي تكمن في اعداد ضخمة من تنوعات البيانات المختلفة الكبيرة، وما هو متوافر من طرق وأساليب إحصائية تقليدية لا يفي بتحليلها بطريقة ملائمة. وبالطبع الحصول على كم كبير من الملاحظات لا يمثل إنجازا في كل الأحوال التي يتطلبها البحث العلمي، بل إن التنوع والاختلاف المرتبط بذلك قد يكون في إطار مستوي الحالة أو الدولة المعينة، كما قد يكون مرغوبا فيه لاستخدام التأثيرات الثابتة أو الطرق الأخرى التي تتحكم في التنوع وعدم الانسجام إلا أنها تقلل أيضا القوة الإحصائية.

٣/٢ توافر البيانات عن أنواع متغيرات جديدة:

صارت كثير من البيانات مسجلة عن الأنشطة التي كان من الصعب ملاحظتها في الماضي. حيث أن كل من بيانات رسائل البريد الإلكتروني، أو بيانات المواقع الجغرافية التي يمكن للأشخاص الحصول عليها حاليا، أو بيانات الشبكات والمواقع الاجتماعية التي تلتقط الارتباطات الشخصية تبرهن جيدا أنها تمثل فرصا سانحة وجيدة للباحثين في مجالات العلوم الاجتماعية المختلفة. وكثير من الباحثين يتفقون على أن الارتباطات الاجتماعية تؤدي دورا مهما في البحث عن وظيفة، أو في تشكيل أفضليات وأوليات المستهلك، أو في نقل المعلومات وتراسلها.

٤/٢ ورود البيانات بهيكلية أقل:

إحدى تداعيات توسع مجال المعلومات المسجلة ترتبط بمجموعات البيانات الجديدة التي تشتمل على هيكلية أقل

وأبعاد أعلي. وفي مثال تجارة التجزئة السابق الإشارة إليه، يتضح أن المعلومات المتوافرة عن المستهلك قد تشتمل على تاريخ التسويق الكامل للمستهلك، الذي من خلاله يمكن إنشاء مجموعة خواص سلوكية على المستوي الفردي للمستهلك والتعرف من خلاله على أبعاد نمط مشترياته. ويتسم هذا البعد بالتحدي حيث ترد ملاحظات المشتريات في شكل متعامد يتضح منه أنواع الملاحظات والمتغيرات وأيهما أقل من الآخر. وتسجل البيانات ببساطة وفقا لتتابع الأحداث بدون تواجد هيكلية أكبر. كما قد يوجد عدد ضخم من الطرق الخاصة بالحركة فيما يتعلق بما هو مسجل في شكل متعامد أو تعاقبي ويسهم في تحديد كيفية تنظيم البيانات غير الهيكلية وتقليل أبعادها وتقييم ما إن كانت الطريقة التي تفرض أوجه الهيكلية لا تمثل الشيء الذي يحصل عليه معظم الباحثين ومتخذي القرارات وحصلواوتمكنوا خبرةمن خلاله علي ميزة فيه، وبذلك يصبح هذا العامل تحديا مشتركا وشائعا جدا في كثير من الدراسات التطبيقية الهادفة.

ومن هذا المنطلق يمكن تطبيق نقطة شبيهة عند التفكير في العلاقات بين سجلات البيانات المتاحة. وتفترض كثير من الطرق وخاصة المستخدمة في المجال الاحصائي التقليدي استقلالية ملاحظات البيانات أو مجموعة منها تسجل في جداول البيانات، أو قد تتصل معا بواسطة الوقت، إلا أنه على سبيل المثال، قد يرتبط الأشخاص في الشبكة الاجتماعية معا في طرق معقدة بدرجة عالية وأن نقطة نمذجتها قد تكون غير مكتشفة بالضبط، مما يستدعي أهمية التعرف على خواص أو أوجه هيكلية اعتماداتها الرئيسية. وبذلك يصبح تطوير الطرق الملائمة للوضاع المختلفة التي يواجها محللو الأعمال يمثل تحديا لكل الباحثين (Imbens et al, ٢٠١١).

٣. البيانات الكبيرة والنمذجة التنبئية:

استخدامات البيانات الكبيرة الأكثر شيوعا بواسطة شركات الأعمال المختلفة التي تختص بتتبع عمليات الأعمال والمخرجات الناجمة منها بهدف بناء ترتيب واسع للنماذج التنبئية Predictive Models. وبينما تكون تحليلات الأعمال عملا رئيسيا وجوهريا لكفاءة وفعالية شركات الأعمال المحسنة، حيث تقع النمذجة التنبئية خلف توافر كثير من المنتجات والخدمات المعلوماتية المقدمة في السنوات الحديثة.

ومن الأمثلة المعينة لذلك والمألوفة للكثيرين ما يرتبط بتوصيات كل من شركة Amazon وشركة NetFixالتين تعتمدانعلى النماذج التنبئية لمشتريات الكتب والأفلام لكلا الشركتين. كما أن نتائج بحث Google وتغذية الاخبار تعتمد على الألجوريثمات التي تتنبأ بمدي توافق صفحات أو مواقع أوقات الويب الخاصة بها. وتحاول وظيفة شركة Apple الآلية بالكامل التنبؤ بباقي نص أو رسالة بريد إلكتروني بناء على أنماط استخدام الشخص المرتبط بذلك في الماضي. كما تعتمد الدعاية والتسويق على الخط على النماذج التنبئية الآلية التي تستهدف الأفراد المحتملين للاستجابة فيما يقدم لهم.

ويمتد تطبيق الألجوريثمات التنبئية إلى ما وراء ما هو متاح على الخط فقط، ففي مجال الرعاية الصحية صار من المألوف حاليا للمأمنين صحيا تكييف وتنظيم مدفوعاتهم وتحديد مقاييس جودة الخدمات الصحية المقدمة المبنية على معدلات المخاطرة المنبثقة من نماذج تكاليف مخرجات الخدمات الصحية الفردية التنبئية. وبذلك يتمثل معدل المخاطرة الفردية في التالي:

مؤشرات الصحة المحددة التي تعرف ما إن كان للفرد أوضاع مرضية مزمنة أم لا، ومع أوزان المؤشرات المختارة المبنية على التحليل الاحصائي يمكن تحديد مدي المخاطرة الفردية، كما تستخدم شركات بطاقات الائتمان نماذج إعادة الدفع والتخلف لتوجيه أنشطة تسعيرتها وتسويقها وتأمينها.

وقامت إحدى شركات منطقة بالو آلتوPalo Altoبولاية كاليفورنيا في الولايات المتحدة وهي شركة Palantis بتطوير ألجوريثمات تستخدم في التعرف على مخاطر الإرهـاب مستخدمة الاتصالات وبيانات أخرى لاكتشاف السوك المخادع في الرعاية الصحية والخدمات المالية المختلفة مما در عليها مكاسب هائلة.

وفي إطار الممارسة الفعلية تعتمد هذه التطبيقات على تحويل كميات كبيرة من البيانات غير الهيكلية في مراتب أوScores تنبئية آلية غالبا ما تكون متدرجة كليا أحيانا في الوقت الحقيقي وبذلك يمكن استخدامها في طرق عديدة منها التالي:

أولا: إمكانية تسريع آلية العمليات الحالية، وفي هذا الصدد تضع شركة Amazon وحدات المطبوعات التي تتنبأ بتوافقها مع المستهلك أو في الوضع المعين، وبذلك تستبدل التوصية التي قد يكون الشخص حصل عليها سابقا من أخصائي أو أمين المكتبة على سبيل المثال.

ثانيا: إمكانية الاستخدام في تقديم مجموعة من الخدمات الجديدة، على سبيل المثال شركة Apple تستقطب كلمة أو جملة مترددة بمعدلات عالية وتقترحها للاكتمال الآلي.

ثالثا وأخيرا: إمكانية الاستخدام لمساندة عمليات اتخاذ القرارات، على سبيل المثال في نطاق محاولات وجهود البنوك في تقليل معالم الخداع والتحايل في بطاقات الائتمان تنفذ سياسة إملاء وفرض الموافقة المسبقة على التصرفات المختلفة، وإقرار أيها يقبل أو يرفض مما يتطلب القيام بدراسة أكثر تعمقا بناء على معدلات التصرف.

إلي جانب ما تقدم من طرق يجب استخدام المراتب أو الدرجات التنبئية في التحليلات المؤداة، حيث يوجد أيضا أعمالا كثيرة تستخدم الأساليب الإحصائية مثل انحدار ريدج Ridge Regression القادر علي تقليل قابليات التغيير وتحسين حدوث نماذج الانحدار الخطي المستقيم، كما أن استخدام أسلوب Lasso الذي يمثل النموذج الخطي الذي يقدر معامل الكثافة والتكاثر يعتبر مفيدا لتفضيل الحلول مع قيم معالم أقل، هذا إلي جانب أساليب تنقيب البيانات ومن ضمنها ألجوريثمات تعلم الآلة Machine Learningفي تحديد تلك الأعمال والتطبيقات التي منها نماذج التصنيف ونماذج التكرارات أو الارتـدادات التي صارت شائعة الاستخدام في مجالي الإحصاء وعلم الحاسب الآلي علي الرغم من ندرة استخدامهما في الدراسات التطبيقية الحالية وخاصة في الواقع المصري. وعلى الرغم من أن وصف الطرق تفصيليا قد يمهد الطريق وراء هذا العمل، إلا أننا نقدم عرضا مختصرا قد يكون مفيدا لتثبيت الأفكار التي تتعلق بالمناقشة اللاحقة.

ومشكلة النمذجة التنبئية يمكن وصفها في تصور مجموعة من المداخل Nالتي ترتبط مع مجموعة متساوية من المقاييس المخرجة N، بالإضافة لمجموعة أقل من المتنبئات Predictors الأساسية K. وفي كثير من الحالات عن المعلومات عن كل مدخل فإنها تكون ثرية وغير هيكلية وعلى ذلك توجد متنبئات كثيرة يمكن انتاجها. وعلى الرغم من ذلك فقد يكون عدد المتنبئات الأساسيةK أكبر من عدد الملاحظات N مما يجعل الاهتمام الواضح يرتبط بالتهيؤ المفرط أو الزائدOverfittingمع أن الواقع غير المفرط يتمثل في المعادلة التالي(K>N) مما يسهم في توضيح وشرح المخرجات الملاحظة بدقة، على الرغم من ضعف الناتج من عينة الأداء.

وفيما يتعلق بهذا الإطـار تتضح الغاية من إنشاء نموذج احصائي يعظم عينة القوة المتنبئ بها، إلا انه في نفس الوقت قد يكون متسما بالتهيؤ المفرط/الزائد Overfitting في الطريقة التي قد تقود لضعف أداء العينة. وتتنوع الطرق المختلفة التي تؤدي للطريقة التي ينشأ فيها النموذج المستخدم للمتنبئات الضرورية غير المفرطة أو الزائدة، على سبيل المثال طريقة Lasso التي يختار فيها معامل القيم المطلقة Absolute Values Coefficient. ومن الشائع والمألوف لتقييم التناوب بين عينة القوة التنبئية والتهيؤ المفرط/الزائد فصل العينة في عينة تدريب مستخدمة لتقدير أبعاد النموذج، وعينة اختبار مستخدمة لتقييم الأداء. ويشار لهذا النوع من النمذجة التنبئية بالتعلم المراقب Supervised Learning. ويوجد أيضا فصل كبير لأساليب البيانات الكبيرة التي يطلق عليها التعلم غير المراقب Unsupervised Learning الذي تتوافر فيه متغيرات المخرج ويهدف لوصف كيف ترتبط مجموعات متغيرات المتنبئ الكبيرة بعضها ببعض. ويشتمل هذا النوع علي طرق مختلفة ميل التجميع Clustering او المكونات الأساسية، ويعتبر مدخلا تصحيحيا متقاطع Cross Validation- قد كون نادر الاستخدام في البحوث التطبيقية المتعلقة بالتخطيط والتنمية الاجتماعية والاقتصادية.

ومـن الافتراضات الأساسية في تعلم الآلـة Machine Learning انها غالبا ما تكون ضمنية مما يتمثل في ثبات البيئة المدروسة نسبيا، بمعني أن عينة التقدير (لكل من عينيتي التدريب والاختبار) تكونا منتجتان بواسطة نفس السحوبات المستقلة التي سوف تنتج مؤخرا العينة المتطلبة للنبؤ. وبالطبع، تبزغ عندئذ البيئات المختلفة عبر الوقت، مما يجعل الافتراض غير متقن. وفي التطبيقات التي تصبح البيانات الجديدة متوافرة لها ومتكررة، يمكن استعادة الألجوريثمات ذاتها باستمرار، وقد يكيف النموذج التنبئي عبر الوقت كلما تغيرت البيئة المحيطة.

وباعتبار كل من المخططين ومتخذي القرارات وراسمي السياسات نفعية طرق تعلم الآلة في تحليلاتهم فإن ما يرد للذهن في الغالب يرتبط بما أشار له روبرت لوكاس Robert Lucas عند رسم سياسة الاقتصاد الشامل الماكرو فيما يتعلق بالنماذج عند محاولة التنبؤ بأثار التغيير على السياسة الاقتصادية علي أساس العلاقات الملاحظة في البيانات التاريخية وخاصة تلك المجمعة بصفة عالية. وعلى ذلك إذا استخدم النموذج التنبئي لتقرير مدي تداخل السياسة فقد لا تكون النتيجة تمثل ما يتنبأ به النموذج لأن تغير السياسة قد يؤثر على السلوك المحدد المنشئ للعلاقات في البيانات، مما يجعل تطوير النماذج تنبئية بدلا من النماذج الهيكلية. وبالطبع، لا يجب نقل هذا النقد الجدالي إلى النموذج التنبئي إن لم يكون ذلك معتمدا على أبعاد كثيرة من الوضع المتاح. على سبيل المثال، من الممكن أن بعض متسوقي شركة Amazon لتسويق المطبوعات لاحظوا كيف ان بعض التوصيات المنتجة المتاحة لهم تغير سلوك تسوقهم لحد ما، على الرغم من أن معظم هذه التوصيات لا يؤدي لذلك التغيير. أي أنه، إذا بدأت شركة Amazon في تقديم خصومات كبيرة على ما تقدمه باستخدام نماذج تنبئية شبيهة، فإنها سوف تستقطب تغيير سلوك كثير من المستهلكين أكثر مما هو متبع حاليا. ويستعرض الجزء التالي أساليب الإحصاء وتعلم الآلة الموظفة في التحليلات والنمذجة التنبئية.

٤. الإحصاء وتعلم الآلة:

وجد مفهوم تعلم الآلة منذ عدة عقود ماضية، وما هو جديد يرتبط بإمكانية تطبيق هذا المفهوم فيما يتعلق بكميات

ضخمة من البيانات. وقد زاد الاهتمام في نظم تعلم الآلة بزوغ وتطور إمكانيات متقدمة لتخزين بيانات أرخص، معالجة موزعة، حاسبات آلية أكبر، وتوافر فرص تحليلات البيانات الكبيرة حاليا.

١/٤ تفسير المصطلحات المستخدمة:

حيث صارت منظمات اليوم تجمع بيانات كبيرة فقد تحولت إلى الاهتمام بما يطلق عليه «علم البيانات Data Science» لاستخلاص ما تتضمنه البيانات من معرفة ومعني لها. ويتضمن ويبني علم البيانات على أساليب ونظريات مستمدة من مجالات تتضمن الإحصاء، تنقيب البيانات، تعلم الآلة والذكاء الاصطناعي، وغيرها. وطبيعة علم البيانات المتعددة التخصص تعني أن هذا العلم تتطلب ممارسته فرق عمل بخبرات في مجالات متنوعة قد يطلق عليهم علماء البيانات.

ويرتكز علم البيانات يعلي «تعلم الآلـة Machine Learning» كفرع من الذكاء الاصطناعي AI الذي يعتمد على الحاسبات الآلية التي تعمل بدون البرمجة الظاهرية لها. والفكرة من آلية بناء النماذج التحليلية التي تستخدم الألجوريثمات التي تتعلم من البيانات التفاعلية المتاحة. وبواسطة اختيار نموذج أحسن، يمكن تحسين النتائج عبر الوقت بتداخل بشري أقل ومحدود لحد كبير. وعندئذ يمكن لهذه النماذج أن تستخدم لإنتاج قرارات مكررة وموثوق منها.

مما تقدم يمكن شرح أسلوب تعلم الآلة المرتكز على إنشاء دراسة النظم التي يمكن أن تتعلم من البيانات لتعظيم وظيفة الأداء مثل تعظيم وظائف منح المكافأة المتوقعة أو حجبها. وتتمثل غاية أسلوب تعلم الآلة في تطوير بصائر متعمقة نابعة من أصول البيانات بطريقة أسرع، واستخلاص المعرفة من البيانات بدقة أعظم، وتقليل المخاطر بقدر الإمكان.

ويجود نوع من التداخل الكبير بين علم الإحصاء وتعلم الآلة، حيث أن كلا من المجالين يرتكز على دراسة التعميمات أو التنبؤات من البيانات. إلا أن هناك اختلافا كبيرا بين كلا المجالين يتمثل في أن الإحصاء يركز أكثرعلى التحليل الاستنباطي أو الاستنتاجي واختبار الفرص لعمل التنبؤات المرتبطة بمجموع أو جمهور أكبر مما تعرضه العينة، إلى جانب ذلك ينظر الإحصاء للأشياء كأبعاد التقديرات، معدلات الخطأ، افتراضات التوزيع وهكذا لفهم البيانات الطبيعية مع مكون عشوائي.

أما أسلوب تعلم الآلة فإنه يستخدم كميات ملاحظات ضخمة ويعتبر فرعا من الذكاء الاصطناعي المرتكز على الآلية المرتبطة بالألجوريثمات التي تتداول الأشياء آليا كتحديد القيم الناقصة، إيجاد التفاعلات وهكذا. ومركزيا لتعلم الآلة تتكون الفكرة التي مع كل تكرار يتعلم الألجوريثم من البيانات المتاحة التي تتداولها، وذلك لقياس ما إن كانت الشركة أو المنظمة المعينة تحسن الأداء أم لا، مما قد يؤدي للنظر في الوظيفة التي ترتبط بالهدف مثل تقليل الفاقد من العمل. ويتفاعل الألجوريثم خلال البيانات حتي مقابلة معيار التقارب، أي استخدام البيانات المحفوظة لرؤية التهيئة المفرطة/ الزائدة Overfitting.

٢/٤ أنواع ألجوريثمات تعلم الآلة:

توجد أربعة أنواع مختلفة لألجوريثمات تعلم الآلة التي يمكن أن تنظم في تصنيف مبني علي مخرج الألجوريثم المرغوب فيه أو على نوع المدخل المتوافر لتدريب الآلة. والمصطلحات المستخدمة في تعلم الآلة تختلف من تلك المستخدمة في الإحصاء. على سبيل المثال، في تعلم الآلة يطلق على الهدف « الموضح Label» بينما في الإحصاء يطلق على الهدف «متغير معتمد Dependent Variable» وعلى ذلك تشتمل أنواع الألجوريثمات الرئيسية لتعلم الآلة علي ما يلي:

التعلم المراقب Supervised Learning

التعلم غير المراقب Unsupervised Learning

التعلم شبه المراقب Semisupervvised Learning

تعلم التقوية/التعزيزReinforcement Learning

والعرض التالي يشرح كل من هذه الأنواع الرئيسية:

١/٢/٤ التعلم المراقب

معظم تعلم الآلة (حوالي ٧٠٪) يكون تعلم مراقب. وعلى ذلك، ألجوريثمات التعلم المراقب تعتبر تدريب يستخدم أمثلة «موضحة Labeled"حيث يكون المخرج المستهدف معروفا. والتعلم المراقب مستخدم بصفة شائعة في التطبيقات التي تستخدم البيانات التاريخية للتنبؤ المحتمل عن الأحداث المستقبلية. على سبيل المثال، يمكن أن يتوقع التعلم المراقب أي معاملات أو تصرفات بطاقات الائتمان المخادعة أو الاحتيالية في بيانات التدريب. ويستلم ألجوريثم التعلم مجموعة مدخلات مع مخرجات صحيحة مرتبطة بها، ويتعلم الألجوريثم بواسطة مقارنة مخرجه الفعلي مع المخرجات الصحيحة، وعل ذلك يمكنه أن يكتشف الأخطاء إن وجدت ويعدل النموذج طبقا لذلك. ويطلق علي المدخلات «الأوجه Features» في تعلم الآلة. وفي حالة الخداع أو التحايل Fraud، فقد تكون أمثلة الأوجه ترتبط بحساب التوازن، عدد المعاملات اليومية، وهكذا. ومن خلال طرق مثل التصنيف، التنبؤ، الانحدار والتباهي المنحدر gradient bosting يستخدم تعلم الآلة المدخلات للتنبؤ بقيم المدخلات الموضحة Labels

وفي هذه الحالة يمكن تطبيق نموذج حالات جديدة لتصنيف المعاملات المخادعة أو المتحايلة Fraudulent ولا يطلق عليها تحديد مرتبة أو درجة Scoring.

٢/٢/٤ التعلم غير المراقب

من حوالي ١٠ أو ٢٠ ٪ من تعلم الآلة يكون غير مراقب، على الرغم من أن هذا المجال ينمو بسرعة. والتعلم غير المراقب هو نوع من تعلم الآلة حيث يشغل النظام على أمثلة بيانات غير موضحة Unlabeled. وفي هذه الحالة، لا يخبر النظام الإجابة الصحيحة. ويحاول الألجوريثم إيجاد الهيكل الخفي أو المتنوع في البيانات غير الموضحة. وبصفة متناقضة مع التعلم المراقب وتعلم التقوية/التعزيز، لا يوجد للأمثلة المدخلة مخرج مستهدف ضمني أو إشارات مكافأة Reward ترتبط بكل مدخل، وغاية التعلم غير المراقب تتمثل في اكتشاف البيانات الهياكل المقدرة طبيعيا أي الأصلية في استخدام الطرق مثل التجميع العنقودي Clustering او تقليل البعد وبذلك يعمل التعليم غير المراقب جيدا على بيانات المعاملات.

وتتنوع كلا من الهياكل الأصلية وطرق التعلم غير المراقب اعتمادا على طبيعة البيانات. على سبيل المثال، البيانات المتواجدة في مشروع فضاء Euclidean التي يمكن أن تكون قد نمذجت هيكليا بواسطة كثافة الاحتمال كما يمكنها استخدام طريقة مثل تجميع عنقودي متوسطات K(K-means clustering).

٣/٢/٤ التعلم شبه المراقب

يستخدم التعلم شبه المراقب لنفس التطبيقات في التعلم المراقب، إلا أن ذلك الأسلوب يستخدم كلا من البيانات الموضحة Labeled والبيانات غير الموضحة للتدريب على أن تكون كمية البيانات الموضحة صغيرة بينما كمية البيانات غير الموضحة تكون كبيرة. ويمكن ان يستخدم هذا النوع من التعلم مع طرق مثل التصنيف، التنبؤ والركود. والتعلم شبه المراقب يعتبر مفيدا عندما تكون التكلفة المرتبطة بالبيانات الموضحة عالية جدا لكي تسمح لعملية التدريب الموضحة بالكامل، إلا أن البيانات غير الموضحة المتزود بها تكون رخيصة. كما قد يفسر التعليم شبه المراقب من خلال طريقتين مختلفتين: في التفسير الأول تستخدم البيانات غير الموضحة Unlabeled لإعلام الكمبيوتر ألجوريثم المعلومات الهيكلية الخاصة بالبيانات المتوافقة مع التعلم المراقب الذي يعتبر الغاية الأصلية. وفي هذه الرؤية، تقدم البيانات غير الموضحة جانب المعلومات المحتاج لها لمساعدة تعزيز التعلم المراقب وخاصة عندما تكون البيانات الموضحة غير كافية. أما في التفسير الثاني فإن الغاية الأصلية

تتمثل في التعلم غير المراقب (التجميع العنقودي على سبيل المثال)حيث ينظر للبيانات الموضحة فيه كجانب معلومات (مؤشرات عنقودية في حالة التجميع العنقودي) لمساعدة الألجوريثم إيجاد هيكل البيانات الأصلية. وفي هذه الحالة، تكون البيانات الموضحة مفيدة بصفة معينة عندما يكون هيكل البيانات الأصلية غير واضح جدا ويطرح تحديات لطرق التعلم غير المراقب. ومن أمثلة ذلك الأولية ما يتضمن تحليل الأشكال Image Analysis (مثل تعريف وجه أحد الأشخاص على الويب) والتحليل النصي Textual Analysis وكتشاف المعرفة Knowledge Discovery.

٤/٢/٤ تعلم التقوية/التعزيز:

مع تعلم التقوية او التعزيز يكتشف الألجوريثم بنفسه أي أفعال تقدم المكافأة الأكبر خلال المحاولة والخطأ. ويستمل تعلم التقوية على ثلاثة مكونات رئيسية هي:

العميل Agent– أي المتعلم أو متخذ القرار،

البيئة Environment– كل شيء يتعامل معه العميل، و

الأفعال Actions– ما يمكن أن يفعله العميل.

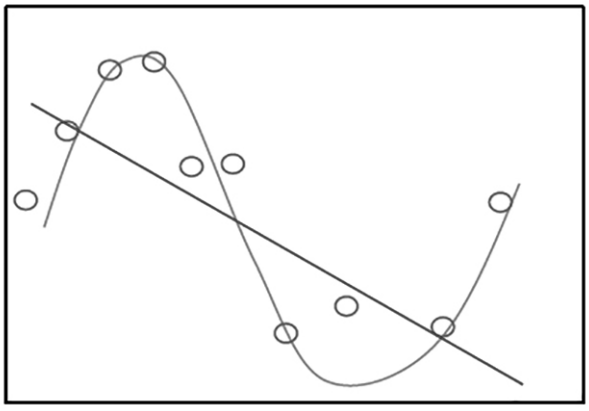
ويهدف هذا النوع من التعلم أن يكون للعميل اختيار الأفعال التي تعظم المكافأة المتوقعة في فترة زمنية معينة. وبذلك وسوف يصل العميل إلي الغاية أسرع جدا بواسطة اتباع سياسة جيدة. وعلي ذلك فإن الغاية في تعلم التقوية تكون لتعلم السياسة الأحسن، كما يستخدم هذا التعلم غالبا في كل من مجالات علم الإنسان الآلي Robotics والابحار Navigation.

ولتعلم التقوية قوة ارتباطات مع الرقابة المثلي، الإحصاءات، وبحوث العمليات، وعمليات قرار ماركوف Markov Decision Processes (MDPs) التي تشكل نماذج شائعة ومستخدمة في تعلم التقوية. وتؤكد عمليات قرار ماركوف (MDPs)أن حالة البيئة تكون ملاحظة تماما بواسطة العميل. وعندما لا يكون الوضع مثل ذلك، يمكن استخدام نموذج عام أكثر يطلق عليه عمليات قرار ماركوف الممكن ملاحظتها لإيجاد السياسة التي تقوم بحل حالة عدم التأكد بينما تعظم المكافأة طويلة الأجل.

٣/٤ التعميم والتقييم واختيار النموذج:

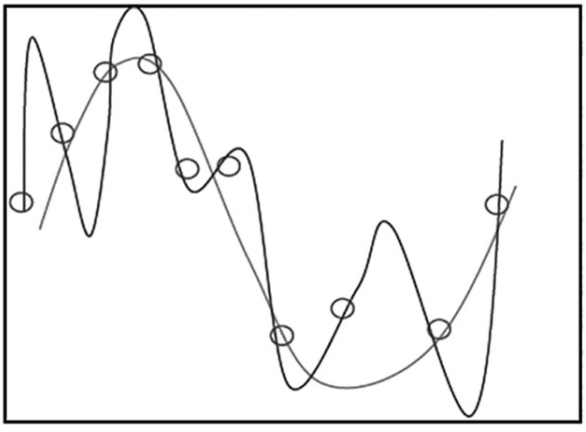
بغض النظر عن الطريقة المستخدمة، كل أنواع تعلم الآلة تطور النماذج التي تساعد آلة التعلم للأداء بدقة على أمثلة أو مهام جديدة او غير مرئية. عندئذ يمكن أن تحسن الآلة هذه النماذج بواسطة التعلم عبر الوقت. ويكون تطوير النموذج الصحيح الملائم مهما جدا ولا يراد أن يكون هذا النموذج كبيرا أو صغيرا جدا ولكن أن يكون صحيحا. واشكل التالي (رقم ١) يوضح مثلا للتهيئة القليلة Underfitting عندما يكون المتنبئ بسيطا جدا لالتقاط الأنماط الملحوظة في البيانات، إلا أنه لا يكون جيدا لحل أمثلة المستقبل. وبذلك يصبح من المفيد الحصول علي نماذج قليلة بحدات

أو الفاظ محدودة جدا إلا أن ذلك النموذج لا يقوم بوظيفة التهيئة جيدا أيضا.



شكل (١) التهيئة القليلة Underfitting

أما اشكل التالي (٢) فإنه يبين التهيئة المفرطة أو الزائدة Overfittingوخاصة عندما يكون المتنبئ به معقدا جدا. وبذلك لا يعمم هذا النموذج أيضا عند محاولة تحديد مرتبة أو درجة Score المجموع أي الجمهور الجديد حيث تكون الرغبة في طلب شيء ما بأبعاد محتملة أقل من استخدام وظائف عقابية أو وظائف التوقف لإيجاد النماذج المهيأة جيدا بدون زيادة أو نقص مع البيانات. وفي الغالب يستخدم علماء البيانات متوسط الخطأ التربيعي أو غير المصنف لبيانات التوقف للقياس عندما يكون النموذج ذا تهيئة زائدة أم لا. وفي هذا الصدد يمكن ملاحظة أن ألجوريثمات تعلم الآلة يمكنها النظر للنموذج وتحديد ما إن كانت المتغيرات تستخدم بكثرة حتي تقوم بتعديل النموذج الخاص بها آليا لاستخدام متغيرات أقل.



شكل (٢) التهيئة المفرطة/الزائدة Overfitting

١/٣/٤ اختيار النموذج:

لبناء نموذج بحجم جيد يحتاج علماء البيانات تقليل درجة التعقيد والحجم الكبير لبيانات ذلك النموذج. وبذلك يحتاج بناء النماذج إلى أهمية التمعن في كل البيانات أو مجموعاتها الفرعية لإنشاء نموذج دقيق لها. وفي هذا الصدد، يوجد أحد ألجوريثمات تعلم الآلة الذي يطلق عليه الغابة العشوائية Random forest الذي صارة أداة تتسم بالقوة في تنقيب البيانات، حيث يأخذ أسلوب أشجار القرار Decision trees الفردية لتجميع البيانات حول محاور رئيسية تتفرع لمحور فرعية حتى تحديد القرار. وعند إدخال مخل جديد للنظام يعمل هذا الألجوريثم او الأداة في تشغيل كل الأشجار حتى تصبح النتيجة إما متوسط أو متوسط مرجح لكل المحاور الطرفية الموصلة.

وعلى ذلك عند مواجهة إمكانية تهيئة تلك الغابة العشوائية من البيانات، يصبح من الضروري بناء أشجار القرار عن مجموعات فرعية عشوائية كثيرة من البيانات وبعدئذ عمل متوسط لها لبناء النموذج النهائي. كما يمكن أيضا تجزئ البيانات إلى متغيرات مختلفة في كل نقطة انشقاق لإنشاء شجرة القرار. وعند توافر مثلا حوالي مائة متغير يمكن النظر في عشرة متغيرات منها فقط عشوائيا فيما يتعلق بكل نقطة انشقاق. وبينما يمكن قياس أشجار قرار فردي من خلال التباين أو التمييز العالي فإن المتوسط المتوصل له يوازي عندئذ المتطرفين Extremes.

وفي هذه الحالة تسمح التكنولوجيات الجديدة مثل التحليلات في الذاكرة لتساؤل البيانات المحفوظة في ذاكرة الوصول العشوائية RAM للحاسب الآلي وعبر البيئة الإلكترونية الموزعة حتى يمكن تجزئ المعالج عبر حاسبات متعددة. ويسمح ذلك السلوب لعلماء البيانات في بناء غابات عشوائية أسرع مما هو متاح.

وفي استخدام نماذج تعلم الآلة في تطبيقات تنقيب بيانات الأعمال، لا يعرف مديري منشآت الأعمال في الغالب الربح او التكلفة المتوقعة من العمل مع عملائهم، إلا أنه عند استخدام محرك أو منقب تنقيب بيانات المنشأة مثل محرك SAS Enterprise Miner الخاص بالنمذجة التنبئية يمكن محاولة اختيار النموذج النماذج التي تعظم الربح أو الإيراد. على سبيل المثال، عند اتخاذ قرار عما يجب أداؤه مع عميل ما، فإن ذلك لا يتعلم بقرار نعم أو لا، بل بدلا من ذلك ترتبط الحاجة بتقرير المخرج المتوقع أو الإيراد المتوقع من اتخاذه القرار، ويعتبر هذا الأسلوب عاملا مهما للإضافة للقرار المطلوب اتخاذه.

٢/٣/٤اختيار النموذج وتقييمه:

بمجرد بناء النموذج يحتاج لتديد مدي صلاحيته حتى يمكن إقرار إمكانيته في القيام بتنبؤات فعالة أم لا. وفي هذا الصدد يمكن استخدام مجموعة بيانات التدريب لتطوير النموذج وبعدئذ تستخدم بيانات معروضة من خارج العينة الأصلية لاختبار هذا النموذج. وعندما لا يتوافر كم وافي من البيانات

التي يسمح لبعضها الحفظ للاختيار، تستخدم عينة فرعية أو عينة بيانات طبقية Stratified، كما يمكن استخدام بعض الأساليب الأخرى التي منها أسلوب K-fold، إلي جانب إمكانية ملاحظة أنه عند تواجد عدد كبير من الملاحظات التي قد تصل لمليون ملاحظة ويعتبر معدل الحدث ١٪، الذي يمكن أ تحديد مدي أفادته لتقييم كل البيانات حتي يمكن تفهم ما إن كان في الإمكان تحديد تصنيفه أو تنبؤ بالحدث. وفي بعض الحالات المعينة، كما في الاحتيال أو الخداع الذي يكون معدل الحدث فيه صغيرا، فإنه يمكن إيجاد عينة زائدة Oversample لتصحيح التحيز في مجموعة البيانات الفرعية وتطوير عينات حيوية تضع ثقلا أكبر إلي معدل الحدث مما يؤدي لنجاح النموذج أحسن.

وقد طورت بعض النماذج لاستخدام قاعدة بيانات التسويق لتحديد فئات أو درجات العملاء. على سبيل المثال، يحتاج السوق لمعرفة أي العملاء الأكثر احتمالا لشراء منتج معين حتى يمكن تقديمه لهم. وفي العادة تشتمل جهود التسويق على معدل حدث صغير يطلق عليه في العادة معدل الاستجابة الذي يكون ١٪ غالبا. وعند تقييم النماذج المستخدمة في قاعدة بيانات التسويق يمكن استخدام الإحصاءات التي تراعي الرفع أو كيفية تأدية النموذج بدرجة تعمق في الملف. وفي هذه الحالة قد لا يكون الشخص مهتما بمعدل التصنيف الخطأ في النموذج، حيث يتاح ١٪ فقط للاستجابة، وفي هذه الحالة يكون نموذج الخمول Null مساوي ٩٩٪. مما تقدم يصبح من الأفضل البدء في تطوير التنبؤات وإنشاء سماتها التي تتعلق بالرفع واختيار النماذج التي تعظم عملية الرفع في الملف.

٥. الفرص التي تتيحها البيانات الكبيرة للسياسات التخطيطية والتنموية:

استخدامات البيانات الكبيرة الأساسية للسياسات التخطيطية والتنموية يوازي تقريبا استخداماتها في شركات ومنظمات القطاع الخاص. ويبدأ العرض التالي بوصف موارد البيانات المتوافرة لدي الأجهزة والمصالح الحكومية، وتحديد كيف أن بيانات شركات القطاع الخاص قد تستخدم في تتبع أنشطتها الاقتصادية والتنبؤ بها بطريقة أحسن. بعدئذ توصف البيانات الكبيرة المستخدمة لإعلام قرارات السياسة أو لتحسين الخدمات الحكومية، وعلى نفس النهج توصف معلومات المنتجات والخدمات المتوصل لها.

١/٥ استخدامات البيانات الإدارية الحكومية:

خلال دور الحكومة في إدارة نظام الضرائب، والبرامج الاجتماعية والتشريعية المختلفة فإنها تجمع كميات ضخمة من البيانات الإدارية المتفرقة. ومن أمثلة ذلك ما تتضمنه مجموعات بيانات المستويات التنفيذية الدنيا الغنية تلك التي تحفظ بواسطة مصلحة الضرائب أو هيئة التأمينات الاجتماعية أوالهيأة العامة للتأمين الصحي علي سبيل المثال لا الحصر، حيث يوجد كم كبير من البيانات غير الموحدة واللامقننة في الواقع العملي، إلي جانب ذلك تنتج المحليات والمحافظات كميات كبيرة من البيانات الإداريــة أيضا وخاصة في مجالات التعليم والضمان الاجتماعي والصحة وغيرها. وتعتبر البيانات الإدارية الحكومية بالتأكيد غير مستقلة بالكامل في كل المصالح والأجهزة الحكومية ويرجع ذلك بسبب الوصول المحدود بل والممنوع الاطلاع عليها من قبل الباحثين والإعلاميين وغيرهم الذين قد يستخدمونها للكشف عن حقائق جديدة تفيد في جهود التخطيط والتنمية. وتتجه مجموعات البيانات الإدارية الرئيسية للحفظ بطرق منفصلة عن بعضها البعض فيما عدا البيانات المتقادمة التي تحدد بالحفظ المستديم فقد حيث قد تنقل لدار الوثائق القومية التابعة لوزارة الثقافة أو لدار المحفوظات العمومية التابعة لوزارة المالية حيث يكون البحث فيها مرتبطا بالتوجهات التاريخية البحتة. وبالطبع يعتبر هذا التوجه المرتبط بعدم إمكانية الحصول علي البيانات الإداريــة الحكومية في البحوث التخطيطية والتنموية مخالفا لما تتبعه كثير من الدول المتقدمة كالدول الأوروبية علي سبيل المثال التي تمتلك البيانات التي تمزج مجموعات البيانات الديموغرافية والتوظيفية والصحية والتعليمية لكل سكانها ومواطنيها، حيث تعتبر البيانات الإدارية الحكومية موردا قويا ومهما في تغطية بيانات سمات الأفراد والكيانات المختلفة عبر الزمن وينشا لها جداول وتحدد أبعاد جودتها مما يجعلها ذات طابع ثري لاستخلاص كثير من المؤشرات ذات الأهمية للتخطيط والتنموية المستهدفة (Card et al, ٢٠١١). إضافة لذلك، تكون قد تكون تغطيتها ذات طابع دولي حيث يمكن ربط مجموعات البيانات الإدارية مع بعضها ببعض، وفي نفس الوقت يمكن الانتقاء منها ما يرتبط بموضوعات معينة.

وفي بعض الحالات التي تسمح المصلحة الحكومية إمكانية الوصول والاطلاع علىمجموعات بياناتها الإدارية قد تتواجد بعض التبعات والنتائج التي ترتبط بالسياسات التخطيطية والتنموية. وفي كثير من الحالات لا يأتي ذلك من خلال أي تصميمات أو مسوح متقنة الأداء، ولكنه يرد من خلال وصف الأنماط الأساسية التي تكمن في مجموعات البيانات المختلفة. على سبيل المثال، استخدم كلا من بيكيتيوسيز(Piketty and Saez, ٢٠٠٣) بيانات نظام استرجاع البيانات IRS لاستنتاج سلسلة المشاركة التاريخية في إيـرادات المواطنينمن خلال تحديد أيهم الأعلى دخلا في الولايات المتحدة الأمريكية، وكان لعملهم هذا تأثيرا كبيرا على مداولات السياسة في الولايات المتحدة وبلورة المناقشات حول عدم المساواة الاقتصادية في دخول المواطنين الأمريكيين.

ومثال آخر قد يختلف في التفاصيل عن المثال السابق إلا انه يشبهه لحد كبير، ويرتبط بما قام بهجون وينبرجJohn Wennberg أستاذ المجتمع وطب الأسرة بمعهد دارتموث للصحة والأداء الكلينيكي Dartmouth Institute of Health and Clinical Practice حيث توصل هو وزملائه من الباحثين إلي تواجد تبيان غير مرغوب فيه في صناعة الرعاية الصحية بالولايات المتحدة الأمريكية، فخلال عقود أربعة (٤٠ عاما) قام بتوثيق التباين الجغرافي في الرعاية الصحية للمرضي في الولايات المتحدة من خلال استخدام عينات كبيرة من بيانات الرعاية الصحية لبيان التباين في نفقات الرعاية الصحية المقدمة للمرضي مما أثر علي مخرجات الصحة المناسبة في الولايات المتحدة، وقد حظي هذا العمل باهتمام واضح من قبل المشرعين الأمريكيين وأدي إلي إصدار قانون «إمكانية منح الرعاية الصحية Affordable Care Act» عام ٢٠٠٩، الذي صار يمثل الدليل الرئيسي لعدم الكفاءة في نظام الرعاية الصحية بالولايات المتحدة الأمريكية.

٢/٥ مقاييس نشاط القطاع الخاص الاقتصادي الجديد:

تقوم الحكومات أيضا بدور مهم في تتبع ومراقبة نشاط القطاع الخاص الاقتصادي. وقد تم عمل هذا التتبع والمراقبة تقليديا من خلال استخدام طرق المسح المختلفة. علي سبيل المثال، يقوم مكتب احصائيات العمل بالولايات المتحدة الأمريكية وكثير من وزارة العمل بالدول المختلفة والجهاز المركزي للتعبئة العامة والاحصاء في حالة مصر أيضا بقياس مستويات التضخم في العمالة بواسطة قيام الماسحين لجمع البيانات عن الأسعار المعلنة والمتوافرة للسلع المتداولة لحوالي (٨٠٠٠٠ وحدة أو سلعة) تختار بعناية فائقة في حالة الولايات المتحدة، حيث تجمع البيانات وتوضع في قوائم تضخم مثل كشاف سعر المستهلك Consumer Price Index (CPI)إلي جانب توفير مقاييس للتوظيف والإسكان ومصروفات المستهلك والأجور التي تعتمد علي طرق مسحية شبيهة.

إلى جانب تلك النوعيات من المسوح توجد مداخل بديلة لجمع البيانات على نطاق واسع حتى ولو كانت بيانات في الوقت الحقيقي مثل البيانات عن الأسعار والتوظيف والإنفاق التي صارت متوافرة على نطاق واسع في كثير من الدول وحاصة المتقدمة منها. على سبيل المثال، يقدم مشروع أسعار المليون وهو مبادرة أكاديمية من قبل أساتذة المعلومات بمعهد ماساتشوست التكنولوجي MITوهما Alberto Carvalho and RoberoRigbon الذيناستخدموا الأسعار المجمعة من مئات تجار التجزئة من خلال مواقع الويب الخاصة بهم على الخط حول العالم في اكثر من ٥٠ دولة، أي إن هذا العمل يمثل مقياس بديل لتضخم سعر سلع تجارة التجزئة، ويعتمد علي بيانات علي أساس يومي لأداء البحوث الاقتصادية والتعرف علي بيانات تضخم التكرار العالي للدول والقطاعات المختلفة. وتستخدم البيانات لإن شاء كشافات الأسعار التي يمكن تحديثها في الوقت الحقيقي. ففي الولايات المتحدة الأمريكية، كما يوجد كشاف ملكية الأعمال الشخصية Business Personal Property (BPP)) الذي يتبع كشاف سعر المستهلك (CPI) نسبيا. وفي بعض الدول الأخرى قد لا تكون مقاييس المسح الحكومية فيها عير موثوق منها أو إنها غير متواجدة كليا. وقد استخدم كافلو(Cavello, ٢٠١٢) بيانات كشاف ملكية الأعمال الشخصية (BPP) لتوثيق أنماط السعر بنفس الطريقة التي استخدمها الباحثون في بيانات كشاف سعر المستهلك ((CPI Klenow and Kryustov), ٢٠٠٨).

وتوجد أيضا إمكانيات شبيهة تتعلق بزيادة قياس إنفاق المستهلك ومعدلات التوظيف. فمثلا يقدم منتج يطلق عليه "SpendingPulse" في سوق بطاقات الائتمان MasterCard بيانات إنفاق العميل في الوقت الحقيقي فيما يتعلق بمجموعات تجارة التجزئة المختلفة، وبذلك تنتج شركة الائتمان MasterCardتقارير دورية تتنبأ بنجاح المخرجات المبنية علي المسح للأمام في طليعة الوقت،

كما أنه بطريقة مشابهة تصدر كل من شركة معالجة البيانات الآلية ADP(وهي شركة مقدمة للسحابة المبنية علي حلول إدارة رأس المال البشري وخدمات تعهيد عملية الأعمال Business Process Outsourcing حيث تقدم خدمات وبرمجيات الأجور والضرائب وغيرها من خدمات إدارة القوي العاملة) وشركة تحليلات مودي Moody's Analytics(وهي شركة تابعة لشركة Moody's Corporation تساعد أسواق رأس المال ومهنيو إدارة المخاطر كما تستجيب لتطور السوق العالمية حيث توفر الأدوات الفريدة والممارسات الأحسن لقياس وإدارة المخاطر من خلال خبراتها في تحليل الائتمان والبحوث الاقتصادية وإدارة المخاطر المالية وبذلك توفر البرمجيات والخدمات الاستشارية لكل ذلك) تقارير شهرية عن توظيف القطاع الخاص المبنية علي بيانات مستمدة من نصف مليون شركة صغيرة ومتوسطة وكبيرة تقريبا التي تقدم شركة ADP برمجيات الأجور لها.

وعلى الرغم من تلك المداخل المرتبطة بالقياسات المختلفة، إلا انها ما زالت تشتمل على بعض القصور النسبي لمقاييس المسح الحكومية، فعلي الرغم من أن عينات البيانات المحددة تعتبر كبيرة، إلا انها تمثل عينات مريحة قد غير متعبة عند عرضها، وبذلك فإنها تعتمد على من يمتلك بطاقات ائتمان معينة مثل بطاقات MasterCard ويقرر استخدامها، أو علي شركات تقدم أدوات وبرمجيات لإدارة سجلات الأجور لعملائها مثل شركة ADP. كما انه من جهة أخري، تكون

البيانات متوافرة وفي تتابع وتكرار وتجزيئ معين، إلى جانب أنه يجدر الإشارة إلي تواجد مسوح كثيرة لا تثير الاهتمامات نتيجة للاستجابات المنتقاة والجودة المرتبطة بالعناصر غير المتجانسة. بالإضافة لكل ما سبق، فإنه من الأفكار الخادعة أيضا، ما يرتبط باستخدام مقاييس غير مباشرة تختص بتساؤلات البحث أو مواقع الوسائط الاجتماعية لتقديم تنبؤات الاحصائيات الاقتصادية المعاصرة على سبيل المثال. وفي هذا الصدد، بين كلا من شوا وفاريان(Choi and Varian, ٢٠١٢) أن محرك بحث Google يمكنه تقديم قياسات دقيقة ترتبط بسلسلة الوقت الاقتصادي تختص بطلبات التوظيف وثقة المستهلك، إلى جانب توضيحهما أيضا أن محرك بحث جوجل يقوم بتحديد نموذج الوقت التي ترتبط بمبيعات السيارات على سبيل المثال، إلى جانب الإمكانية في تحسين متوط الخطأ التنبئي الجذري من خلال إضافة مقياس اتجاهات جوجل للبحث المعاصر.

وعلي الرغم من التقاط كلا من شوا وفاريان(Choi and Varian, ٢٠١٢) تمكنا من التقاط قليل من سلاسل الوقت الاقتصادي المعين، إلا أن المدخل الذي اتبع يمكن تطبيقه لسلاسل البيانات عن انفاق المستهلك وغير ذلك من الأنشطة المختلفة. وبالطبع، تتمثل إحدى التحديات الممكن مواجهتها في تواجد فئات عديدة أو آلاف تساؤلات البحث المختلفة التي قد تتنبأ بالإنفاق في فئات أو تجمعات مختلفة. ولذلك اقتر كلا من اسكوتوفاريان(Scott and Varian, ٢٠١٣) أهمية تطبيق الباحثين مدخلا آليا متضمنا أدوات التعلم الإحصائية السابق التعرض لها، وتحليلات سلع المستهلك الكثيرة أو الضيقة إلى جانب سلسلة الوقت القصير.

في هذا الإطار يمكن الشك في شيوع وانتشار أنواع الكشافات أو مؤشرات الوقت الحقيقي للنشاط الاقتصادي أو غيره. إضافة لذلك، كما قد تنشئ اتجاهات جوجل كشافا معينا باستخدام المعلومات النابعة من تساؤلات البحث في محركه. إلي جانب ذلك، نلاحظ أنه في الوقت الحالي ينشر موقع التواصل الاجتماعي تويترTwitter كشافا يوميا مبني علي سياق الرسائل المحملة عليه. وبذلك فقد نستنتج من العرض السابق أنه حتى من خلال البحث الشاق قد يصعب التعرف على بيانات تفصيلية ترتبط بالتوظيف، إقراض المستهلك، نفقات الائتمان او التسويق يوميا سواء في الواقع أو تلك المحملة على الإنترنت، ويدعونا ذلك لتصور كيف أن أنواع البيانات ذات التكرار العالي يمكن أن تكمل أو تحل محل سلاسل البيانات التقليدية عن الأنشطة المختلفة التي ترتبط بحياة ومقدرات المواطنين.

٣/٥ تحسين عمليات وخدمات المصالح والأجهزة الحكومية:

من التغييرات الكبيرة في الأعمال المعاصرة ما يتمثل في التداول والقرارات الوتينية بواسطة كميات تحليلات البيانات الكبيرة، التي قد تكون في بعض الشركات القليلة من خلال التجارب التي تتم لذلك (Varian, ٢٠١٠). وحالياعلىالرغم من أن بعض المصالح الحكومية وخاصة في بعض الدول الأجنبية المتقدمة تتسم بالذكاء بصفة متزايدة وخاصة فيما يتصل باستخدام تحليلات البيانات لتحسين عملياتها أو خدماتها، إلا أن معظم المصالح الحكومية وخاصة في الدول النامية ومن ضمنها مصر فبالتأكيد تعتبر متخلفة لحد كبير عن شركات القطاع الخاص الأحسن بصفة خاصة في استخدام تحليلات البيانات الكبيرة الناجمة من بياناتها الإداريـة الضخمة، كما انها تواجه كم كبير من التحديات وخاصة فيما يتعلق ببنياتها الأساسية واحتياجات القوي العاملة بها. على سبيل المثال، وصف تقرير مجموعة دراسة JASON الصادر عام ٢٠٠٨ بعض التحديات فيما يتعلق بالقوات المسلحة الأمريكية التي يجب أن تعالج وتحلل كميات بيانات الاستشعار الضخمة التي صارت متوافرة والبيانات المستخلصة من الاتصالات وطلعات البيانات (Jason, December Study Group ٢٠٠٨).

وفي بعض الأحيان، تجمع المصالح والأجهزة الحكومية كم كبير من البيانات التي قد تكون مفيدة في توجيه قرارات السياسات العامة إلا انها لسوء الحظ لا توظف بفعالية. على سبيل المثال، هيئة مثل الهيئة العامة لتأمين الصحي بها سجلات بيانات عن كل حالات علاج المؤمنين صحيا من موظفي الدولة وبذلك تمتلك كم ضخم من مجموعات البيانات الصحية التي تخص الحالات المرضية التي ترددت علي مستشفياتها، إلا أن هذه البيانات لم تحلل بطريقة علمية تسمح بالتعرف علي التكلفة والعائد المفصلة للعلاجات المقدمة والاجراءات المتخذة، وللتنبؤ باحتياجات الهيئة في التخطيط المستقبلي لها من حيث الرعاية الصحية للمرضي وعلاجاتهم.

وعلي ذلك يجب علي الهيئات والمصالح الحكومية المختلفة اكتشاف مجموعات بياناتها وإتاحة الفرصة للوصول إليها من قبل الباحثين والمخططين وراسمي السياسة سواء بالمنظمة أو من خارجها لكي يحللوها بهدف تحسين عمليات الهيئة أو المصلحة ذاتها ورسم أبعاد خططها المستقبلية علي أسس علمية مستمدة من مؤشرات تحليلات بياناتها. وأح أمثلة المحليات ما يرتبط بمدينة نيويورك في الولايات المتحدة الي صارت تقدم وتوفر كتالوج شامل بمجموعات البيانات المتوافرة والممكن أنزالها من موقعها علي الإنترنت NYC Open Data الذي يشتمل علي بيانات جغرافية، وبيانات عن الأماكن والمدارس والمستشفيات، إلي جانب بيانات الإنفاق العام، ومعلومات عن الطرق والنقل والمترو،

واستهلاك الكهرباء، واحصاءات الجرائم ومئات أنواع البيانات الأخرى. وقد استخدم هو (Ho, ٢٠١٢) هذا المصدر المفتوح والمتاح لتحليل مدي تفتيش أبعاد الصحة في مطاعم مدينة نيويورك ووثق الأبعاد الصحية المختلفة لنوعيات المطاعم بالمدينة، وقد استخلص من تحليلاته بوجود اهتمام قليل جدا فيما يتعلق بالتفتيش الصحي عليها ويشكل ذلك كم كبير من المشكلات الخطيرة فيما يتعلق بصحة المترددين عليها وتحديد درجة ومرتبة المطعم ذاته. إلي جانب ذلك أخذت الحكومة الاتحادية (الفيدرالية) بالولايات المتحدة علي عاتقها القيام بممارسة شبيهة لما قامت به مدينة نيويورك في إعداد موقع علي الإنترنت http://www.data.gov. الذي وفر آلاف مجموعات البيانات الحكومية واتاحتها للمواطنين المهتمين في البحث والتحليل والإعلام.

٤/٥ معلومات المنتجات والخدمات:

معظم تطبيقات بيانات شركات القطاع الخاص السابق مناقشتها تستخدم النمذجة التنبئية فيما يتعلق بآلية عمليات أعمالها، أو لتحسين أو تطوير منتجاتها أو خدماتها الجديدة. وبينما تكون الأجهزة والمصالح الحكومية مشغولة أيضا بما تقدمه من خدمات، إلا انها غير ملمة في الغالب بكثير من الأمثلة التي توضح تلك الأنشطة، على الرغم أن مجموعات البيانات الحكومية المتاحة قد تكون نابعة فيما يتعلق بإنشاء أنواع منتجات المعلومات الشائعة أيضا في القطاع الخاص. ومن الأنشطة الحكومية الممكن تصورها فيما يتصل بالمنتجات ما يرتبط بحماية المستهلك، وفي هذا النطاق يكمن التحدي في حماية المستهلك في جعل الأفراد بعيدين عن اتخاذ قرارات يصلون لها من خلال إمكانية التنبؤ ثم يتأسفون عنها بعدئذ. مع العلم بأن الاقتصاديات السلوكية أكدت أن أحد طرق التعامل مع هذا النوع من التوازن يكون من خلال عمل إطار القرارات (على سبيل المثال القرارات المختارة جيدا)، أو من خلال عرض المعلومات بعناية فائقة. وفي هذا الإطار يمكن للأفراد الانتهاء من اتخاذ قرارات مالية روتينية قد ترتبط بشراء شقة، الادخار للتقاعد، تخطيط ابعاد الرعاية الصحية، الخ. بدون توفر معلومات جيدة عن النتائج والتوابع المالية اللاحقة. وبذلك تعتبر أنواع النماذج التنبئية المختلفة جيدة لإنشاء تلخيص المعلومات الشخصية المحتاج لها. وكم مستهلك يقتنص الفرص السانحة أمامه فيما يرتبط بالوضعية المالية التي تحدد؟ وما مدي الرسوم المدفوعة من قبل مستهلك لمنتج أو خدمة مالية؟ وما التكلفة المتوقعة للمرضي الذين يختارون نوعا من العلاج المعين؟ وبينما لا تكون الحكومة هي الكيان الصحيح لإنشاء هذه الأدوات، فإن المعلومات التي تجمعها تكون مفيدة بالتأكيد.

ومن الأفكار الخلافية لحد كبير ما يرتبط باستخدام النمذجة التنبئية لتحسين الخدمات الحكومية المستهدفة. على سبيل المثال، قد يكون في الإمكان تصور الجدال النفعي بأن الرعاية الصحية يجب أن تسجل الأفراد بناء على استجابتهم المحتملة للعلاج وتغطي أنواع العلاج عندما لا يتعدى حدا أو مستوي معين. وشبها لذلك، برنامج حسم الضرائب الذي يستهدف تقديم مزايا اقتصادية معينة للمولين عند سداد مستحقاتهم الضريبية المتأخرة قد يكون الكثر فعالية عندما يستهدف للعائلات مثلا.

تعتبر الأمثلة السابق طرحها مفيدة لأنها تتصل بأنواع الأشياء التي تقوم الشركات والهيئات المختلفة بتحليلات بياناتها المتعلقة بما تقدمه من منتجات أو خدمات كل الوقت ترتبط بعملائها ومستهلكيها وكيفية جذبهم وحثهم للإقبال عليها من خلال تقديم خصومات ومزايا معينة، وموافقاتها على طلبات التأمين والإقراض عند تلبيتهم معايير محددة.

٦. الفرص التي تقدمها ثورة البيانات الكبيرة على السياسات التخطيطية والتنموية:

فيما يتعلق بالسؤال عن كيف تؤثر ثورة البيانات المعاصرة على البحوث المرتبطة أساسا برسم السياسات التخطيطية والتنموية المعاصرة وصولا لجودة النتائج المستخلصة؟ إلي جانب تحديد جودة الطرق المستخدمة في تحليلاتها وكيفية تدريب المخططين وراسمي السياسات والاقتصاديين وغيرهم؟ للإجابة على تلك التساؤلات المرتبطة بمؤثرات ثورة البيانات الكبيرة وتحليلاتها نلاحظ أولا أن التأثير الأكثر وضوحا هو ذلك الذي يسمح لقياس الآثار والمخرجات من نتائج متوصل لها بطريقة أحسن من تلك المستخدمة للطرق التقليدية في ذلك، حيث يمكن للبيانات الشاملة تحديد وطرح أنواعا جديدة من الأسئلة كما تساعد في تصميمات البحوث الجديدة التي تسهم في إبراز نتائج وتداعيات السياسات والأحداث والمتغيرات على الأبعاد التخطيطية والتنموية المختلفة. ومن الإمكانيات الأقل وضوحا بالمؤثرات المتوقعة ما يرتبط بالمصالح والأجهزة الحكومية العديدة والمتنوعة التي تجمع كم كبير ومتنامي من البيانات المفيدة لتوجيه قرارات السياسات المستهدفة، إلا انها لا توظف بفعالية وكفاءة مرجوة. على سبيل المثال، يمكن اعتبار ما ان كانت جهود التخطيط والتنمية متمكنة من الإفادة في توظيف بعض أساليب تنقيب البيانات على تحليلاتها المختلفة أم لا، ولماذا يكون ذلك الجانب أقل وضوحا في رسم السياسات المختلفة؟ هذا القصور الواضح في توظيف الأساليب الآلية المتقدمة ما زال هو السائد في تحليلات البحوث القائمة حتى الآن. إلي جانب تلك الحقيقة المتمثلة في قصور أساليب التحليلات المتقدمة يري كثير من الباحثين تواجد تمييز واضح بين النمذجة التنبئية وما تتضمنه من تحليلات تنبئية وأساليب الاستدلال السببية، ونتيجة لمداخل التعلم الاحصائي الراهن تقتصر المساهمة المتوقعة. وعلي ذلك يمكن التفكير الجدي

في تعليم واكساب الاقتصاديين والباحثين مهارات وكفاءات توظيف الأساليب التحليلية وخاصة التنبئية المتقدمة المرتبطة بتنقيب البيانات للوصول لذكاء الأعمال فيما يرتبط بمؤشرات رسم السياسات التخطيطية والتنموية. ويستعرض العرض التالي كل من طرق القياس الجديدة المرتبطة بالتصميمات البحثية ومعالم تأهيل وتنمية القوي العاملة المتطلعة بذلك.

١/٦ طرق القياسات الجديدة المرتبطة بالتصميمات البحثية:

تعتبر مجموعات البيانات الإدارية الكبيرة النطاق وبيانات شركات ومنشآت اقطاع الخاص والقطاع العام الجديدة المتدفقة بمعدلات عالية جدا مهمة وجوهرية في مساعدة أي تنوع في تصميمات البحوث الجديدة وخاصة التطبيقية منها. ومن الأمثلة البارزة في هذا الاتجاه الدراسة التي قام بها كل من شيتي وفريدمان وروكوف (Chetty, Friedman and Rockoff, ٢٠١١) عن تأثيرات المدرسين طويلة الأمد وقيمهم المضافة علي مخرجات الطلاب عند الكبر، أي توقع الآثار المترتبة علي ضرورة توافر مدرسين أحسن علي العملية التعليمية في المدي البعيد. وقد تضمنت هذه الدراسة بيانات إدارية عن حوالي مليونين ونصف مليون تلميذ ملتحقين في مدارس مدينة نيويورك الأمريكية وتحديد التوقعات من المكاسب التي يحصلون عليها عند الكبر بعد عشرين عاما، وكان التساؤل الرئيسي للدراسة ما إن كان الطلاب الذين حصلوا على قيمة مضافة أعلي في المدي القصير قد حصلوا أيضا على مكاسب أعلي ترتبط بالتبعية للكبار أم لا؟ وقد قيست القيمة المضافة للمدرسين بواسطة مراتب ودراجات Scores اختبار صمم لذلك، الذي استنتج عددا من النتائج المثيرة للانتباه التي يمكن استخلاص فوائد عديدة منها تتعلق بتحليلات البيانات الإدارية ذات النطاق الكبير كما يالي:

أولا: القدرة علي ربط بيانات نتائج الاختبار القيمة الضافة للمدرسين وسجلات الضرائب اللاحقة لعدد كبير من الطلاب في الكبر بعد عشرين عاما، وهذا النوع من القياس والمضاهاة لكم كبير من البيانات الإدارية الكبيرة قد يعتبر صعبا أو حتى مستحيلا مع بيانات تجميعية أو عينة عشوائية صغيرة.

ثانيا: طبيعة بيانات الضرائب الطويلة الأجل تجعل في الإمكان التقاط كلا من مكاسب الطلاب الكبار والمعلومات عن أباءهم في الفترة التي يعتمدون فيها عليهم.

ثالثاوأخيرا: طبيعة بيانات درجات أو مراتب الاختبار التجزيئية تسمح بفحص الافتراض الرئيسي المحتاج له للتعريف أن الطلاب لا يوزعون على المدرسين وفقا لمقدرتهم المحدودة.

ومن الأمثلة الأخرى الحديثة التي تستخدم كلا من البيانات الإدارية ذات النطاق الكبير بالإضافة لبيانات ملكية شركات القطاع الخاص ما يرتبط بما قامت به شركة أوريجون للرعاية الصحية Oregon's Medicaidمن تجربة تتصل بالرعاية الصحية المقدمة منها وما توصل له من نتائج في السنة الأولي من التجربة (Finkelstein et al, ٢٠١٢)، حيث أنه في عام ٢٠٠٨ استخدمت الشركة يناصيب Lottery لاختيار مجموعة من الأفراد المؤهلين للانضمام للتأمين الصحي التي تقده الشركة. وقد أدي هذا اليناصيب تجربة طبيعية كبيرة وفرصة سانحة لدراسة آثار تقديم تأمين صحي سخي مقدما مزايا للملتحقين به. وقد جمع الباحثون نتائج اليناصيب وبيانات القيد اللاحقة مع السجلات الإدارية فيما يرتبط بالوفيات وما قدمته المستشفيات من علاجات مع بطاقات الائتمان المحصل عليها المرتبطة ببيانات المسح. وكانت النتائج المتوصل لها من هذه التجربة مثيرة للاهتمام في بعد العام الأول حيث أن الأفراد الين تم تغطيتهم للحصول علي الرعاية الصحية Medicaid تمتعوا برعاية صحية لاحقة أكبر، وائتمان طبي أقل، والحصول علي تقارير صحية ذاتية أحسن (علي الرغم من ان الدراسة اللاحقة لنفس المؤلفين وجدت دليل تحسين اقل من تنوع مصادر القياس الحيوية Biometrics). كما وضحت التجربة أيضا نفس الفوائد طويلة المدي لمجموعات البيانات الأخرى، حيث تمكن الباحثون من أخذ مجموعة فرعية من سكان أو جمهور ولاية أوريجون لاحقة وتحديد مدي تخصيصات العلاج المقدمة لهم من قبل مستشفيات الولاية من خلال سجلات المرضي، بالإضافة لتواريخ الائتمان المتمتعين به في مجموعات البيانات الشاملة، وقد سمح لهم هذا النهج في تتبع تداعيات ونتائج تجربة الرعاية الصحية Medicaid من خلال استخدام عدد من المقاييس المختلفة.

والمثال الثالث يرتبط باستخدام المقاييس لتحديد مؤشرات الخدمات المقدمة من الشركات لعملائها ومستهلكيها، وخاصة فيما يخص تجارة الإنترنت أي التجارة الإلكترونية علي الإنترنت باستخدام بيانات الملكية ذات المدى الكبير التي يتحصل عليها من خلال التعاون مع خدمة eBay التي ترتبط بتجارة تجزئة السلع والخدمات علي شبكة الويب، ما بينه إيناف وآخرون (Einav, Knoepfle et al, ٢٠١٤)في دراستهم التي استخدموا فيها تصفح مفصل لبيانات شراء المستهلكين المستخدمين لموقع eBay (الذي يتردد عليه أكثر من ١٠٠ مليون مستهلك في الولايات المتحدة فقط) لتحديدمدي تأثير ضرائب المبيعات علي المشتريات المتاحة علي الخط أي من خلال الإنترنت فقط عندما يكون البائع أي تاجر التجزئة في نفس ولاية المشتري، كما قد لا يحصلون ضرائب المبيعات علي المشتريات التي بين الولايات التي يمكن حسابها لجزء كبير من تجارة الإنترنت. وقد وجد أن البيانات المجمعة من تدفقات المبيعات من ولاية لأخري تقدم تقديرات مرنة على

ضرائب المبيعات، على الرغم من أن تصفح البيانات المفصلة عن تلك المبيعات أدي للحصول على أدلة لمستوي متدني في تحصيل ضرائب المبيعات المتطلبة.وفي هذا الصدد، أمكن إيجاد مجموعة أفراد يدخلون على موقع eBay لتصفح ورؤية نفس الوحدة أو السلعة المعينة، إلا ان بعضهم إذا كان متواجدا في نفس الولاية كبائع يكون خاضعا للضريبة، أما من لا يتواجد في نفس الولاية فلا يخضع للضريبة. عندئذ يمكن مقارنة نزعات مشتريات المجموعتين من البائعين من خلال التعرف على آلاف الوحدات وملايين جلسات تصفحها. ومن خلال هذه الدراسة أمكن التوصل لاستجابات ضريبية مهمة وتحديد أدلة إحلال منتجات شبيهة بديلة (لكنها غير خاضعة للضرائب) على الرغم من الاستجابة الأقل جدا المتوقعة من تحميلات سعر تجزئة السلعة.

وفي دراستين أخرتين (Einav,Farronto et al, ٢٠١٣ &Einav, Kuchler et al, ٢٠١٣)أمكن دراسة التسعير واستراتيجيات البيع علي الخط باستخدام تصميم بحث مختلف يتمتع بميزة طبيعة البيانات المتاحة علي شبكة الإنترنت الجزئية، وقد أمكن أخذ مجموعة الوحدات الموضوعة سنويا علي موقع eBayومئات الآلاف من الوحدات المعرفة التي وضعت كمبيعات مرات متعددة بواسطة نفس البائع' إما بطريقة متزامنة أو من خلال التتابع مع التسعير أو الرسوم او آليات المبيعات المختلفة. وبهذا الأسلوب يصبح استخدام التجارب البحثية لمبيعات البائعين لتقدير درجة تشتت ترتبط بالعمر، والمنتجات المتبقية، وكيفية استجابة المستهلكين ما يشحن لهم بطريقة تتسم بعدم الشفافية المرتبطة بأسعار الشحن على سبيل المثال. ويمكن ان نستنتج من ذلك أن بحوث تجارة الإنترنت أو التجارة الإلكترونية تحددمدب الإفادة من البيانات الجزئية فيما يتعلق بالقيام بالتجارب الطبيعية او البيانات التي ترتبط بالمستهلكين الأفراد والوحدات المباعة حيث يمكن التمتع بميزة توافر التفاصيل المؤسسية المعينة، أو ما يخص مستوي التنوع الدقيق الصعب الاكتشاف من بيانات تجميعية أكبر.وكما في الحالات السابقة من دراسات التي تعتمد على البيانات لإدارية الكبيرة الحجم، تتوافر فرصا أخرى للحصول على بيانات غنية أي ثرية عن الأفراد المدروسين (على سبيل المثال عند تجزئ المستهلكين بواسطة تواريخ الشراء)، أو لاكتشاف تنوع في التبعات والنتائج المستخلصة من تجربة معينة، على سبيل المثال إحلال وحدات مختلفة في حالة تغير السعر.

ومن الأوجه الأخرى المرتبطة باستخدام المقاييس في التحليلات في مثل تلك البحوث التي تشتمل على تقديرات مبنية على تجارب صغيرة كثيرة، من المؤكد ان بعض أجزاء التجارب يعاني من مشكلات عديدة حيث يصعب تأكيد المصداقية المتطلبة لكل تجربة صغيرة بانفراد. على أي حال، من مزايا البيانات الكبيرة الحجم والمجال ما يسمح للاستراتيجيات من تنشيط وتقوية النتائج والافتراضات الرئيسية. فقد استخدم إينافوكوشلروآخرون (Einav,Kuchler et al, ٢٠١٣) تعاريف بديلة لمجموعات وحدات عديدة حتى يتأكدوا من التعريف الأصلي بأنه غير عريض، كما درسوا موضوع الضرائب الموصوفة من قبل، حيث فحصوا النشاط اللاحق الخاص بالمستخدم في نفس جلسة التصفح مما يمكن من إعادة تأكيد التفسير النسبي للنتائج المتوصل لها.

ويلاحظ مما تقدم من دراسات اعتماد الشركات بكثافة على تحليلات البيانات الكبيرة المتدفقة والمرتبطة بعملياتها اليومية، أنه أصبح من الأسهل والأكثر فعالية لها القيام بالتجارب المختصة بتحليلاتها لأن ذلك سوف يجعل أولامن الأسهل جدا إجراء عملية البيع عندما يكون التسعير أو الآليات الأخرى آلي الطابع؛ ثانياعندما تكون استراتيجيات التسعير منفصلة وجزئية للشركات يصبح اجراء التجربة أسهل وأقل ملاحظة وغير محفوف بالمخاطر، وفي الحقيقة، كثير من المنصات المتواجدة على الخط كجزء من عمليات الشركة تستخدم مشاركة عملياتها الصغيرة كمنصة التجربة، وبمجرد التقاط البيانات بسرعة يصبح من الأسهل والأرخص التقاط نتائج التجربة وخاصة عندما تنجز بنجاح؛ وأخيرا مع الاستراتيجيات الآلية يصبح ممكنا للشركات استخدام استراتيجيات متعددة في نفس الوقت، وفي بعض الأحيان عندما تكون عشوائية فيما يتعلق بمجموعة العملاء المقدم لهم خيارا أو أكثر يوجد أيضا نوعا من الأهمية حتى لبعض التجارب العشوائية الظاهرية.

٢/٦ تأهيل وتنمية المخططين والباحثين في السياسات التخطيطية والتنموية:

مما سبق استعراضه من دراسات اعتمدت على استخدام البيانات الكبيرة وتحليلاتها وقياساتها يتضح أهمية هذا النهج في عند اتخاذ قرارات السياسات التخطيطية والتنموية لصالح المنظمات والدول على حد سواء. وفي هذا الصدد قد تستخدم التحليلات المتطلبة في المداخل الفكرية والطرق الإحصائية المألوفة للمخططين وراسمي السياسات التي تمثل العلاقات بين المتغيرات المختلفة مثل كيف يمكن الحصول على علاجات معينة، الالتحاق بالمدرس في المراحل التعليمية المختلفة، نشر فرص التأمين الصحي لكل المواطنين، زيادة الحصيلة الضريبية من خلال تعميم حوافز الخصومات الضريبية الخ. وكثير من الدراسات إن لم يكن معظمها في الاقتصاد الشامل الماكرو تتضمن هذا النوع من الهيكلية المرتبط بعلاقات التغاير المزدوجة Bi -variant التي لا تمثل العامل السببي غالبا، وتساوي المجموع الكلي مع الجزء المنجز بواسطة الرقابة على المتغيرات الأخرى.

وفي مقابل ذلك، فإن مدخل النمذجة التنبئية التي تم التطرق

لها في هذا العمل يكون ذو أبعاد متعددة. والتركيز لا ينصب على كيف أن البعدالواحد يؤثر على قياس نتيجة أو مخرج ما، ولكن كيف أن النتيجة أو المخرج تتغير مع عدد كبير من المتنبئات الجوهرية. وفي هذه الحالة قد يستخدم المحلل أو لا يستخدم نظرية السببية لأن المتنبئات تتوافق معها. من هذا الاختلاف النظري يمكن طرح السؤال التالي هل أساليب البيانات الكبيرة المألوفة في الإحصاءات التقليدية سوف تكون مفيدة في البحوث التخطيطية والتنموية؟

من المحتمل أن تكون إجابة ذلك التساؤل بنعم أي إنها إيجابية، حيث أن من التطبيقات التي تم اكتشافها في الواقع من خلال عدة دراسات كثيرة مثل دراستي (Belloni et al, ٢٠١٢ &Belloni, Chernozhukov and Hansen, ٢٠١٢) بينت أن أساليب تعلم الآلة السابق التعرض لها بالتفصيل في هذا العمل تستخدم لتحسين كفاءة معالجة آثار الدراسات عندما تكون الدراسة إما كبيرة تتضمن عددا كبير من المتغيرات المتواجدة بالفعل أو البديلة، إلا أنه عند استخدام الركود أو الانحدار العقابي إما لتعريف مجموعة من أساليب الرقابة الأمثل أو مجموعة تجارب ذات أعداد كبيرة تعتبر أحسن عند الاستخدام.

ومن استخدامات النمذجة التنبئية الأساسية ما يرتبط بتباين الشركات والمنظمات فيما يتعلق بالنماذج والتحليلات التخطيطية والتنموية المستخدمة للإحصاءات. ففي إحدى الدراسات عن الائتمان وأسواق التأمين (Finkelstein et al, ٢٠١٣) تم اكتشاف أن أسلوب المراتب أو الدرجات Scores عند تحديد مخاطر التأمين الممكن التنبؤ بها التي تلخص التباينات والاختلافات بين الأفراد تعتبر قليلة نسبيا، على الرغم من ذلك الأسلوب يكون إما مفيدا جدا في الأنشطة التأمينية سواء كان ذلك للأفراد المختارين بصفة دقيقة للتغطية التأمينية، أو لتكييف الأسعار التي تقدر في السوق للتكلفة المرتبطة بالاختلافات المحتملة المقدرة للأفراد المأمنين.

وفي تلك الأمثلة من الدراسات السابقة، يتضح أن الباحثين المخططين والمرتبطين بأبعاد التنمية يعتبرون مستهلكين لنماذج تعلم الآلة، إلا أنهم ليسوا منتجيها. وبذلك، يمكن تصور تطبيقات إضافية يكون فيها المخططون مهتمون بتوصيف التباين بين الأفراد، المنتجات، الخدمات أو الشركات عندما يحللوا الاختلافات في القرارات أو الأثار المترتبة عليها. وفي مثل هذه الحالات، يمكن أن تقدم أساليب تعلم الآلة طريقة مفيدة للحصول على ملخصات كميات كبيرة من المعلومات الإحصائية ذات البعد الواحد عن الكيان المدروس مثل معدل ملخصات ائتمان المستهلكين الذي يمثل في تاريخ الاقتراض وإعادة المدفوعات التي غيرت الهيكلية في تلخيص المخاطر العادية. ومن النقاط المرتبطة بذلك المراتب أو الدرجات التنبئية التي يمكن أن تكون مهمة جدا لدراسة ما يختص بها. على سبيل المثال، معدلات مخاطر الصحة تقدم خريطة للديموغرافيات الفردية والرعاية الصحية المستخدمة في التنبؤ بها في المستقبل. وبذلك يمكن طرح سؤال مهم يحظى باهتمام كبير هل هذه العلاقات ثابتة عند ما يتواجد تغيير في البيئة المحيطة؟ فمثلا عند الجهات أو الشركات المؤمنة التي تقرر تحصيل مدفوعات إضافية أعلي، فإن العلاقات الديموغرافية السابقة وما سبق تطبيقه في الماضي والوضع الحالي لن يبقي كما كان بالنسبة للمأمنين، مما يمثل قضية تحتاج لحل سريع. ومن خلال استخدام أساليب النمذجة التنبئية التي تحتاج للتقدير والتقييم على أساس كل حالة لفهم الحدود المتصلة بكيف تكوت العلاقات المتنبئ بها بعيدا عن العينة المختارة صحيحة، ومتي تغير السياسة هذه العلاقات.

٣/٦ احتضان التباين والاختلاف:

من الممارسات الشائعة في الدراسات التخطيطية والتنموية التطبيقية ما يتمثل في تصاميمها التي تساعد تعريف متوسط التأثير لسياسة معينة. وفي الغالب يكون الباحثون لتلك الدراسات ملمين جيدا أن الوحدات المعالجة من الأفراد، المجموعات، الشركات، أو المنتجات تكون متباينة ومختلفة لحج ما، وعلى ذلك فمن المحتمل ألا تتضمن السياسة على آثار موحدة.ففي الغالب تتطلب حدود البيانات تقديم متوسط تلك الآثار، وحتى مع المستوي الدقيق الميكرو فغن التركيز يكون على المتوسط بعد تقدير فردي يكون أسهل في العرض أو الاستخدام عند تشكيل تنبؤات سياسة من خارج العينة.

ومبدئيا يمكن أن تسمح بيانات النطاق الكبير مع الخصائص الفردية الغنية في تقديرات آثار السياسات المختلفة المحددة بطريقة أحسن. وفي هذه الحالات قد يتصور الفرد أن بعض الدراسات قد تتحول من قياس وتقرير متوسط التقدير، تجاه بناء الأداة المطلوب توظيفها، حيث أن الغرض من النموذج القياسي لالتقاط الآثار المترتبة على القرار المعين أو التنبؤات السياسية يرتبط أساسا بالمجاميع الفرعية المختلفة الكبيرة. ومثل ذلك، إمكانية اعتبار مشكلة كتاب دراسي ترتبط بتعريف تعظيم الربح العائد من سعر منتجات الشركة، حيث أن من مداخل المعايير المستخدمة في المنظمة الصناعية يرتبط بالحصول على بيانات عن المبيعات في أسعار مختلفة، ومحاولة عزل التنوع في الأسعار التي تعرف مدي استجابتها للطلب بوضوح، ويستخدم هذا النوع لتقدير منحنى الطلب الذي يواجه الشركة. ومرونة منحنى الطلب يترجم في السعر الأمثل الممكن ان تتبناه تكاليف الشركة.

مما تقدم يمكن اقتراح أن للشركة بيانات عن عملائها الذين يمكن تصنيفهم في مدي واسع من الطرق بحيث يمكن وضع أسعار تمييزية يمكن الاستجابة لها. وفي هذه الحالة، قد لا يرغب الباحث تقدير مرونة فردية، بل بدلا من ذلك يطور

ألجوريثم يقوم بتصنيف المستهلكين في أنواع عديدة، ويقدر مرونة الطلب والأسعار المثلى التي تفضل لكل نوع. هذا النوع من المخرجات التطبيقية صار مألوفا لفترة ما لكثير من الشركات، فعلي سبيل المثال تقوم شركات التأمين تفصيل ما يقدموه لعملائهم. وتقدم البيانات الكبيرة ما يجعل هذا التحليل ممكنا ومقبولا في قطاعات أخري من الاقتصاد القومي، علي سبيل المثال محلات البقالة أو السوبرماركت للبيع بالتجزئة تقدم في الوقت الحالي خصومات معينة علي أسعار بعض السلع التي تسوقها. وبينما بعض الأمثلة التي ذكرت ترتبط بسياسات التسعير مثلا، فإن هناك كثير من الأنشطة الأخرى التي تطبق سياسات شبيهة أخرى كثيرة كما فيما يرتبط بحوافز ومكافآت المقدمة من شركات وأجهزة التأمين والرعاية الصحية، الضرائب، التعليم وغيرها. ففي حالة التعليم على سبيل المثال يعتمد التحفيز علي المرحلة الدراسية، حجم الفصل الدراسي، مزيج الطلاب والبنية الأساسية المتوافرة المرتبطة بالموقع المميز، والمدرسين المؤهلين، والتغذية المقدمة، والتكنولوجيا المتوافرة وغيرها. وكمية البيانات الأساسية المجمعة قد تعتمد علي كل مجموعات تلك الأبعاد وغيرها من بيانات الجمهور المستهدف، وبذلك يمكن استخدام أساليب التحليلات التنبئية التي تساعد في رسم السياسة واتخاذ القرار المناسب.

٧. التحديات:

توجد تحديات كثيرة تواجه المخططين ومتخذي قرارات السياسات المستخدمة الذين يواجهون بكم كبير من مجموعات البيانات الكبيرة والمتجددة باستمرار ويأملون في الاستفادة القصوى منها. وتشتمل هذه التحديات علي كسب إمكانية الوصول لهذه البيانات، تطوير إدارة وبرمجة البيانات المحتاج العمل معها، مع التفكير في المداخل الابتكارية لتلخيص ووصف وتحليل المعلومات المتضمنة في هذه البيانات الكبيرة.

وفي إطارتحديات الوصول للبيانات، فإن الدراسات المتعلقة بموضوعات مثل اقتصاديات العمل، الإنتاجية، والاستهلاك العائلي تعتمد تقليديا على بيانات المسوح التي تقوم بها الأجهزة الحكومية في العادة كما في حالات التعداد السكاني. وتوجد لكثير من هذه البيانات بروتوكولات منشأة لكيفية الوصول لها واستخدامها. وفي بعض الحالات كما في حالة بيانات التعداد السكاني قد تكون هذه البروتوكولات معقدة من المحتمل ألا تشجع عدد كبير من الباحثين، إلا أنها على الرغم من ذلك تعكس جهدا ظاهرا وملموسا إما للوصول والاستخدام أو الحجب للسرية التي تتصف بها من وجهة النظر السياسية المطلقة.

وما زالت كثير من النظم المتاحة حاليا تختص بالبيانات الإداريـــة ذات النطاق يمكن استخدامها في الـدراسـات التخطيطية والتنموية مثل تلك المتعلقة بالرعاية الصحية، التأمينات الاجتماعية، الدخول، الربط الضريبي، الخ. وقضايا الخصوصية التي ترتبط بكميات البيانات الكبيرة تعتبر مهمة أيضا. وفي هذا الصدد أشار كارد وآخرون (Card et al, ٢٠١٠) أن كثيرا من الدول الأوروبية مثل النرويج، السويد، والدنمارك خطت خطوات واسعة في اتاحة الوصول للبيانات المختلفة لتسهيل إجراء البحوث المعتمدة عليها، وخبرة هذه الدول تقترح أن الوصول الأعرض والأعم يكون ممكنا كما أن وضع بعض القيود القليلة للوصول إلى البيانات يمكن أن يكون له تأثير مهم على كمية البحوث وجودتها لحد كبير.

ومن الملاحظ أن كثيرا من البيانات الجديدة التي تم مناقشتها تغطي الشركات الخاصة، والوصول إليها نشئ قضايا عديدة للباحثين، فيما يلي:

أولا وأكثر وضوحا لا تريد كل شركة العمل مع الباحثين، بينما تري بعض الشركات الأخرى إن ذلك قد يعود بالنفع والإفادة لها. وعلى الرغم أن ذلك يعتبر طريقة مفيدة للتعلم من الباحثين من خارج الشركة، يري البعض الآخر أن تلك الطريقة قد تتضمن نوعا من الإرهاق على الشركة وقد يركز على مخاطر تؤثر سلبا على برامج الدعاية لها.

والباحثون المتعاونون مع الشركات يوقعون على اتفاقيات وعقود تمنع الإفصاح عن المعلومات السرية التي تحددها الشركة، كما قد يواجهون بعض التحفظات على بعض الأسئلة المطروحة في الدراسة. والخبرة المستخلصة من ذلك توضح أن فؤاد العمل مع بيانات الشركة ترجح التكاليف بصفة كبيرة، إلا أن ذلك يتطلب جهدا مضاعفا من قبل الباحث والشركة لتطوير نوع من التعاون الناجح بينهما.إلي جانب ذلك، يمكن أن تكون بيانات شركات القطاع الخاص محدودة بطرق معينة، فغالبيتها تشتمل علي معلومات عن عملائها فحسب. وهؤلاء العملاء لا يعتبروا ممثلين حتى مع قطاع من الصناعات المعينة، كما أن كثيرا من مجموعات بيانات شركات القطاع الخاص تجمع لأغراض المعاملات والتصرفات فحسب، ونتيجة لذلك تشتمل على مجموعة معلومات معينة تعتبر مثالية لبعض الأغراض ولكن لبس لكل الأغـراض. على سبيل المثال، السجل الإلكتروني لزيارة الطبيب للمستشفى قد يسجل فقط بيانات الحالات المعالجةالتي كشف عليها، إلا انه لا يسجل بالضرورة أي نوع من المعلومات الصحية عن الحالات المعالجة كالقياسات الببليومترية للمريض وغيرها من الأعراض والعلاجات المقدمة التي قد تكون متاحة في سجلات متنوعة ومنفصلة عن المرضي. نفس الشيء بالنسبة لسجل معاينة مأمور الضرائب لبيانات الممولين قد يرتبط ببيانات عن سجل ضريبة الممول وتحديده للحصيلة الضريبية المتطلبة منه ولكنه لا يرتبط بباقي السجلات الخاصة بالمورد في الجهات الأخرى التي تحدد أبعاد دخله والايراد المتحصل

عليه منها وكل ذلك يمثل تحديات تواجه محللي بيانات الأفراد والشركات.

وفيما يتعلق بتحديات إدارة البيانات واستخدام الحوسبة، فمن الطرق التي وصفت من قبل بعض المعلقين فيما يتعلق بالبيانات الكبيرة تطلب استثمارا في الوقت والجهد والموارد من قبل الإدارة. وبصفة افتراضية تستثمر معظم شركات الإنترنت الناجحة والشركات ذات البيانات الكبيرة في قطاعات الاقتصاد الوطني الأخرىفي أنشطة ترتبط بتخزين البيانات ومعالجة البيانات الموزعة وغيرها، كما تستثمر أيضا في تعين متخصصي ومهندسي الحاسبات المهرة. ومن الملاحظ أن تلك الشركات عند تأجير خدمات اخصائيو أو خبراء البيانات بغية تحليل البيانات للحصول على الأنماط التطبيقية، فإنها تبحث بصفة عامة على الأفراد المدربين في علم الكمبيوتر بدلا من المخططين والاقتصاديين والاحصائيين للقيام بذلك. وهذا يحدد أن مستقبل المخططين والاقتصاديين والاحصائيين الراغبين العمل مع مجموعات البيانات الكبيرة يتطلب منهم التعرف على الأقل لبعض أدوات علم الكمبيوتر وخاصة ما يرتبط بلغة التساؤل الهيكلية SQLألجوريثمات تعلم الآلة المعيارية التي سبق استعراضها في هذا العمل حتى يمكنهم تجميع أطر عمل التخطيط والاقتصاد النظرية مع القدرة في تطبيق الأفكار الفعلية بكفاءة وسرعة فيما يتصل بالتعامل مع البيانات الكبيرة.

وفيما يتعلق بتحديات طرح الأسئلة الصحيحة، تتمثل بعض الملاحظات الإضافية بالارتباط والعمل مع مجموعات بيانات كبيرة غنية جدا التي لا تكون هامشية أو ثانوية لاكتشاف الأسئلة الصحيحة منها الممكن الإجابة عليها بطريقة مقننة. وفي الماضي القريب، كان الباحث في مقدرته فتح ملفات البيانات الخاصة به علي الكمبيوتر الشخصي ويحصل علي الأوجه الرئيسية للأسئلة والاجابات المحتاج لها، إلا أنه مع مجموعات البيانات الكبيرة فإن ذلك يتطلب جهدا مضاعفا للقيام بالمهام المدركة، مثل استخلاص وتلخيص بدائل مختلفة واكتشاف العلاقات فيما بينها. وفي الوقت الحالي يمكن ملاحظة أن بعض أطروحات الدكتوراه صارت تستخدم ظاهرة البيانات المتاحة علي منصات تجار التجزئة التي يضمها موقع eBay المتاح علي شبكة الويب للوصول لمؤشرات ونتائج مستهدفة من خلال التساؤلات المستقرأة من تحليلات تلك البيانات الكبيرة. نفس النتائج يمكن التوصل لها من البيانات الكبيرة المتاحة عليمنصات الائتمان في إطار موقع Persper.com، أو علي منصات الترويح والسفريات المتاحة علي موقع .Airbnb com وغير ذلك من منصات الإدارة المالية. وقد تحولت كثير من هذه المشروعات بطريقة ناجحة جدا في اكتشاف وتحديد ما هو متواجد من بيانات وإدارتها بتحليلات تسهم في الوصول للنتائج المستهدفة. وقد مثل ذلك إطارا خصبا لتحديد مزايا وعيوب البيانات المتاحة، وتطوير الاستراتيجيات والطرق المفيدة لتنظيم البيانات واكتشاف الأسئلة المختلفة. وعلى ذلك فإن الاختلافات بين بحوث الماضي المرتبطة بدراسة عينات قليلة مع بحوث المستقبل مع مجموعات البيانات الكبيرة التي زاد الوصول لها صار يشغل الباحثين كثيرا في الوقت الحالي وسوف يكون له مردودا علي السياسات التخطيطية والتنموية للمنظمات والدول علي حد سواء.

٨. الخلاصة:

مما سبق من عرض يتضح وجود قليل من الشك في أنه في السنوات المقبلة سوف تغير ظاهرة البيانات الكبيرة والتحليلات التنبئية وصولا لمؤشرات ونتائج المستقبل من النظرة المحدودة المرتبطة بعينات البيانات المحدودة المتعامل معها في كثير من البحوث الحالية. وبالتأكيد سوف يتطلب ذلك استخدام طرق وأساليب متقدمة تعتمد علي علوم الحاسب الآلي كاستخدام لغة التساؤلات الهيكلية ألجوريثمات لغة الآلة وغيرها من أساليب تنقيب البيانات لاكتشاف المعرفة المتطلبة.

وقد ظهر حديثا مصطلح «علم البيانات» الـذي صار يشير لمجال علمي نامي مهتم بجمع كميات بيانات كبيرة وإعدادها وتحليلها وإدارتها وعرضها. وعلى الرغم من أن هـذا المصطلح يرتبط بقوة مع مجالات مثل قواعد البيانات وعلم الحاسب الآلي وتنقيب البيانات والاحصاء إلا أنه يتضمن أنواعا من المهارات المختلفة المحتاج لها، ويشتمل علم البيانات على تحليلات البيانات كمكون أساسي له. كما أن هذا العلم الحديث يمثل مجموعة مبادئ رئيسية تساند وتوجه استخلاص المعلومات والمعرفة من البيانات وبذلك يعتبر المجال والمفهوم الأكثر ارتباطا به هو «تنقيب البيانات» الذي يمثل استخلاص المعرفة الفعلي من البيانات عبر التكنولوجيات المتضمنة لتلك المبادئ، وفي هذا الصدد توجد مئات ألجوريثمات تنقيب البيانات المختلفة مع مدي كبير من التفاصيل لطرق المجال، وعلى الرغم من ذلك فإن علم البيانات يتضمن أكثر من ألجوريثمات تنقيب البيانات. حيث انه على القائمين به من علماء البيانات ومحللي البيانات القدرة على رؤية مشكلات الأعمال من منظور البيانات. كما يوجد هيكل أساسي لتفكير تحليل البيانات، والمبادئ الأساسية المطلوب فهمها. كما ان طرق ومنهجية التكنولوجيا الحديثة تعتبر حيوية لهذا العلم، إلى جانب تواجد مجالات معينة مثل الحدس، الابتكارية، والمعرفة لتطبيقات معينة التي يجب أن تكون واضحة بالنسبة لهم، وفي نفس الوقت يقدم منظور علم البيانات للممارسين هيكلية وقواعد لمعالجة مشكلات استخراج المعرفة المفيدة من البيانات الكبيرة المتاحة.

وفي هذا الإطار وظفت ثورة البيانات وعلم البيانات الحديث للتنمية المستدامة التي تعمل علي تحويل الطريقة التي تؤدي بها الحكومة والمواطنين والشركات الأعمال، حيث أنها

تعرف بالانفجار الحادث حاليا في توافر موارد البيانات والتكنولوجيات الحديثة السريعة التطور والنمو، إلى جانب تكلفة أدوات جمع البيانات الرخيصة التي تتراوح من المصدر الضخم للبيانات إلى الأشكال الملتقطة بواسطة الأقمار الصناعية التي غيرت جميعها الطريقة التي تؤدي بها الأعمال وعملت علي زيادة توافر البيانات للكل. وقد حدي ذلك بأن مجموعة الخبراء الدولية عن ثورة البيانات للتنمية المستدامة (IEAG) للأمم المتحدة في اجتماعها عام ٢٠١٤ إلى إلقاء الضوء على الفرص والتحديات التي يواجها العالم في تحسين البيانات للتنمية المستدامة.

وفي إطار ما سبق إثارته عن ثورة البيانات الحالية يصبح من المؤكد أن التحول الحالي في استخدام البيانات وتحليلاتها وخاصة التنبئية سوف ينتشر ويوطد دعائمه كما أن إبداعات التغيير سوف تأخذ مجالا أوسع حيث توجد حاجة ملحة لها مع جعل ما هو حادث مفهوما ومقبولا مما سوف يغير الفكر الإداري في إدارة الأعمال كليا.

المراجع:

-1. Belloni, Alexander et al (2012). "Sparse models and methods for optimal instruments with an application to eminent domain." Econometrics, Vol. 80, No. 6, pp. 2369-2429.

-2. Belloni, Alexander, Chernoehukov, V. and Hanssen, C. (2012). "Inference on treatment effects after selection amongst high-dimensional controls." London: Centre for Microdata Methods and Practice [working Paper No. CWP10/12]

-3. Card, D. et al (2011). Expanding access to administrative data for research in United States. Arlington, VA: National Science Foundation Directorate of Social Behavior and Economic Science [NSF SBE 2020 White Papers].

-4. Cavallo, A. (2012). "Scraped data and sticky prices." Cambridge, MA: MIT [Sloan Working Paper].

-5. Chetty, R., Friedman, J. and Rockoff, J. (2011). "The long-term impacts of teachers: Teachers value-added and student outcomes in adulthood." Cambridge, MA: National Bureau of Economic Research [MBER Working Paper No. 17699].

-6. Choi, H. and Varian, H. (2012). "Predicting the present with Google trends." Economic Record, No. 88, pp. 2-9.

-7. Einav, Liran, Farronto etal (2013). "Selection or moral hazard in health insurance." American Economic Review, Vol. 103, No. 1, pp. 178-219.

-8. Einav, Liran, Kochler, et al (2013). "Learning from seller experiments in online markets." Canbridge, MA: National Bureau of Economic Research [NBER Working Paper No. 17385].

-9. Einav, Liran, Knoefle, D. et al (2014). "Sales taxes and Internet commerce." American Economic Review, Vol. 104, No. 1, pp. 1-24.

-10. Finkelstein, Amy et al (2012). "The Oregon health insurance experiment: Evidence from the first year." Quarterly Journal of Economics, Vol. 127, No. 3, 1057-1106.

-11. Hastie, T. et al (2008). The elements of statistical learning: Data mining, inference and prediction. New York: Springer-Verlag.

-12. Imbens, G. et al (2011). Clustering, spatial correlation and randomization inference. Cambridge, MA; Harvard University Memo

-13. Jason Study Group (December 2008). "Data analysis challenges: JSR-08-142." [http://www.fas.org/irp/agency/dod/Jason/data.pdf].

-14. Klenow, P. J. and Kryvstov, O. (2008). "State-dependent or time-dependent pricing: Does it matter for recent US inflation?" Quarterly Journal of Economics, vol. 123, pp. 863-904.

-15. Piketty, T. and Saez, E. (2003). "Income inequality in the United States, 1913-1998" Quarterly Journal of Economics, Vol. 118, No. 1, pp. 1-39.

-16. Scott, J. and Varian, H. (2013). "Bayesian variable selection for now casting economic time series." San Diego, CA: ASSA Annual Meeting (Presentation ppt.).

-17. Varian, H. (2010). "Computer-mediated transaction." American Economic Review Papers and Proceedings, Vol. 100, No.2, pp. 1-10.