استخدام القدرات التحليلية للبيانات الضخمة في صنع التقارير المتولدة ذاتيا

طالب الماجستير : غيفار محمد

ماجستير علوم الويب – الجامعة الافتراضية السورية

إشراف الدكتور: وسيم الجنيدي

الملخص

مع التزايد الكبير في البيانات الهائلة المنتشرة على مواقع الويب أصبحت عملية التقاط هذه البيانات وتحليلها واستخلاص النتائج الإحصائية منها أمرا ضروريا ومفيدا بالنسبة للشركات العاملة في هذا المجال من أجل خلق بيئة تنافسية وابتكار أفكار جديدة تساعد في تحقيق متطلبات الزبون وبالتالي تحقيق الربح.

يهدف هذا البحث إلى استخدام أدوات تحليل البيانات من أجل توليد تقارير تحليليه وإحصائية يمكن أن تساعد وتدعم متخذ القرار والشركات في تعزيز القدرة الديناميكية للتلاؤم مع بيئة العمل المتطور، واستخدام خوارزميتن أساسيتن للتنبؤ بالسلاسل الزمنية (XGBOOT – Auto Regressive) والمقارنة بينهما ومن ثم دمج نتيجة الخوارزمية الأولى في الخوارزمية الثانية ومناقشة النتائج.

الكلمات المفتاحية: تحليل البيانات الضخمة – التعليم الآلي – التنقيب في البيانات – القدرات الديناميكية – البيئة غير المستقرة.

Big Data Analytics Capabilities for Self-Reported Data

Abstract

With the huge increase in the huge data spread on websites, the process of capturing this data, analyzing it and extracting statistical results from it has become a necessary and useful matter for companies working in this field in order to create a competitive environment and invent new ideas that help achieve customer requirements and thus achieve profit.

This research aims to use data analysis tools in order to generate analytical and statistical reports that can help and support decision-makers and companies in enhancing the dynamic ability to adapt to the evolving work environment, and the use of two basic algorithms for forecasting time series (XGBOOT - Auto Regressive) and comparison between them and then integrating the result of the first algorithm in the second algorithm and discuss the results.

Key words: Big Data Analytics – Machine Learning – data mining - Dynamic capabilities - environmental uncertainty.

1- المقدمة:

يزدهر "عصر البيانات" حاليًا ، حيث يتم إنتاج بيانات جديدة من جميع الصناعات والهيئات العامة بمعدل غير مسبوق. أدت هذه الظاهرة إلى ضجة كبيرة ، حيث تسعى المؤسسات جاهدة للاستفادة من تحليلات البيانات الضخمة من أجل خلق قيمة [1] . نتيجة لذلك ، هناك اهتمام كبير من الأكاديميين والباحثين على حد سواء بالقيمة التي يمكن أن تخلقها المؤسسات من خلال استخدام تحليلات البيانات الضخمة [2] . بعد التوسع السريع في حجم البيانات وسرعتها وتنوعها ، تم توثيق تطورات جوهرية من حيث التقنيات اللازمة لتخزين البيانات وتحليلها وتصورها. ومع ذلك ، هناك القليل من الأبحاث حول كيفية حاجة المؤسسات إلى التغيير لاحتضان هذه الابتكارات ، وما القيمة التجارية التي يمكن الحصول عليها من [3].لا يزال البحث التجريبي حول قيمة تحليلات البيانات الضخمة في حالة بدائية ، وهو أمر مثير للدهشة ، نظرًا لطفرة الشركات التي تستثمر في البيانات الضخمة. جاءت معظم التقارير حول القيمة التجارية للبيانات الضخمة حتى الآن من الشركات الاستشارية والصحافة الشعبية ودراسات الحالة الفردية التي تفتقر إلى الرؤية النظرية. ونتيجة لذلك ، هناك فهم محدود لكيفية تعامل الشركات مع البيانات الضخمة الخاصة بها ، والدعم التجريبي نادرًا ما يدعم الادعاء بأن هذه الاستثمارات تؤدي إلى أي قيمة تجارية قابلة للقياس [4].

تعتبر معالجة هذه الفجوات الحرجة أمرًا مهمًا نظرًا لوجود القليل من المعرفة حول كيفية الاستفادة من تحليلات البيانات الضخمة على مستوى الشركة ، ومن خلال الآليات التي يمكن إنشاء القيمة بها. في هذه الدراسة ، نبني على فكرة القدرة على تحليل البيانات الضخمةBDAC ،تفترض هذه الدراسة أن البيانات الضخمة هي مورد ضروري ، ولكنها ليست شرطًا كافيًا لإحداث مكسب لقيمة العمل. من أجل التمكن من الاستفادة من البيانات الضخمة لدعم وتوجيه عملية صنع القرار الاستراتيجي ، من الضروري وجود

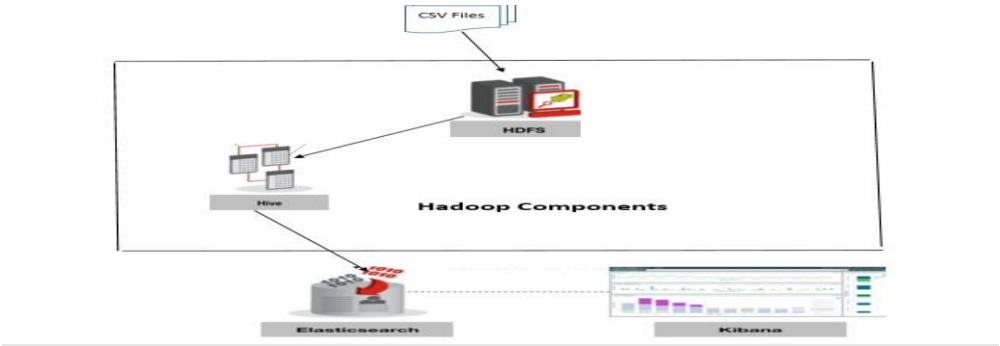
عدد من الموارد التكميلية ، والتي تعمل بشكل متواز على دفعBDAC بشكل عام للشركات. على هذا النحو ، يجب على الشركات أن تكتسب وتطور مزيجًا من الموارد التكنولوجية والبشرية والمالية وغير الملموسة لإنشاءBDAC التي يصعب تقليدها ونقلها. على الرغم من بعض الدراسات النادرة التي تفحص البيانات الضخمة من خلال هذا المنظور الشامل [5] لا يزال هناك فهم تجريبي محدود للآليات التي يمكن من خلالها لـ BDAC توليد قيمة تجارية. نتج عن ندرة العمل في هذا الاتجاه عدم فهم القيمة المحتملة لتحليلات البيانات الضخمة ، وترك المجربين في وضع غير مؤكد عند مواجهة مثل هذه التطبيقات في شركاتهم. للحصول على أي آثار نظرية وعملية ذات مغزى ، وكذلك لتحديد المجالات الحاسمة للبحث في المستقبل ، من المهم فهم كيفية تشكيل المكونات الأساسية لتحليلات البيانات الضخمة وكيف تؤدي إلى قيمة العمل [1] .

2- الهدف من البحث:

يهدف هذا البحث إلى استخدام أدوات تحليل البيانات من أجل توليد تقارير تحليليه وإحصائية يمكن أن تساعد وتدعم متخذ القرار والشركات في تعزيز القدرة الديناميكية للتلاؤم مع بيئة العمل المتطور من خلال دراسة العلاقة بين قدرة تحليل البيانات الكبيرة ومدى تأثيرها على ابتكار أفكار جديدة بنوعيها (الجزرية -الإضافية).

استخدام خوارزميتن أساسيتين للتنبؤ بالسلاسل الزمنية (XGBOOST – Auto Regressive) حيث يتم استعراض نتائج الخوارزمية الأولى ومن ثم تحسين النتيجة عن طريق دمج نتائجها مع الخوارزمية الثانية.

ما يهدف الباحث لبناؤه:



الشكل يبين المخطط التفصيلي لسير عملية تحليل البيانات.

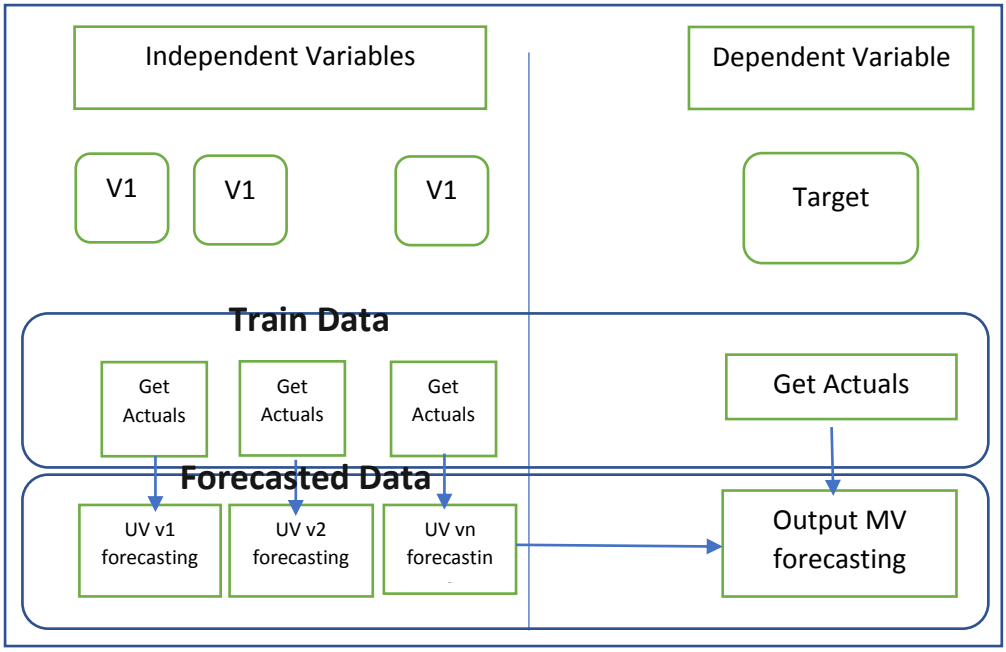
طريقة العمل:

1- تحميل ملفات CSVوجعلها متاحة وقابلة للمعاجلة من قبل Hive

2- تحميل البيانات إلى Elastic search باستخدام Map-Reduce

3- عرض البيانات وتحليلها باستخدام KIBANA

4- التنبؤ بالنتائج بالاعتماد على خوارزميات التنبؤ للسلاسل الزمنية (Auto Regressive - XGBoost)



الشكل يبين: مخطط التنبؤ بالبيانات

3- مواد وطرق البحث:

أساليب تحليل البيانات الضخمة:

تحتاج البيانات الضخمة إلى تقنيات غير عادية لمعالجة حجم كبير من البيانات بكفاءة خلال أوقات تشغيل محدودة .يتم تشغيل تقنيات البيانات الضخمة بواسطة تطبيقات محددة. تتضمن تقنيات البيانات الضخمة عددًا من التخصصات، بما في ذلك الإحصاء، واستخراج البيانات، وتعلم الآلة، والشبكات العصبية، وتحليل الشبكات الاجتماعية، ومعالجة الإشارات، والتعرف على الأنماط، وطرق التحسين، وأساليب التصور. هناك العديد من التقنيات المحددة في هذه التخصصات، وتتداخل مع بعضها البعض كل ساعة [6]

اعتمد الباحث في بحثه على الطريقتين التاليتين:

- **التنقيب في البيانات:(Data mining) استخراج** البيانات هو مجموعة من التقنيات لاستخراج المعلومات القيمة (الأنماط) من البيانات، بما في ذلك تحليل المجموعات والتصنيف والانحدار وتعلم قواعد الارتباط .أنها تنطوي على أساليب تعلم الآلة والإحصاءات .يعد التنقيب عن البيانات الضخمة أكثر صعوبة مقارنة بخوارزميات التنقيب عن البيانات التقليدية .إذا أخذنا التجميع كمثال، فإن الطريقة الطبيعية لتجميع البيانات الضخمة تتمثل في تمديد الأساليب الحالية (مثل التجميع الهرمي و K-Mean و Fuzzy CMean)حتى يتمكنوا من التعامل مع أعباء العمل الضخمة [7] تعتمد معظم الإضافات عادة على تحليل كمية معينة من عينات البيانات الكبيرة، وتختلف في كيفية استخدام النتائج المستندة إلى العينة لاشتقاق قسم للبيانات الإجمالية.
- **تعلم الآلة (Machine learning)** : تعلم الآلة هو موضوع مهم للذكاء الصنعي يهدف إلى تصميم خوارزميات تسمح لأجهزة الكمبيوتر بتطوير

سلوكيات تستند إلى بيانات تجريبية .السمة الأكثر وضوحا لتعلم الآلة هي اكتشاف المعرفة واتخاذ القرارات الذكية تلقائيا .عندما يتعلق الأمر بالبيانات الضخمة، نحتاج إلى توسيع نطاق خوارزميات تعلم الآلة، كل من التعلم الخاضع للإشراف والتعلم غير الخاضع للإشراف، للتعامل معها. أصبح التعلم العميق بالآلات بمثابة واجهة بحثية جديدة في مجال الذكاء الصنعي [8].

سلسلة القيمة للبيانات الضخمة:

تم استخدام سلاسل القيمة كأداة لدعم القرار لنمذجة سلسلة الأنشطة التي تؤديها المنظمة من أجل تقديم منتج أو خدمة ذات قيمة إلى السوق [9] .

تحدد هذه السلسلة الأنشطة الرئيسية عالية المستوى التالية

- **الحصول على البيانات (Data Acquisition)** :هي عملية جمع وتصفية وتنظيف البيانات قبل وضعها في مستودع بيانات أو أي وسيلة تخزين أخرى يمكن إجراء تحليل البيانات عليها.
- **تحليل البيانات (Data Analysis)**: تهتم هذه المرحلة بجعل البيانات الخام المكتسبة القابلة للاستخدام في صنع القرار وكذلك للاستخدام في المجال المحدد.
- **معالجة البيانات (data Curation)** :هي الإدارة الفعالة للبيانات على مدار دورة حياتها للتأكد من أنها تقابل متطلبات جودة البيانات اللازمة للاستخدام الفعال.
- **تخزين البيانات (Data Storage)** :هو الاستمرارية وإدارة البيانات بطريقة قابلة للتطوير التي تلبي احتياجات التطبيقات التي تتطلب الوصول السريع إلى البيانات.

- **استخدام البيانات (Data Usage)** :يغطي أنشطة الأعمال التي تعتمد على البيانات والتي تحتاج إلى الوصول إلى البيانات وتحليلها والأدوات اللازمة لدمج تحليل البيانات في نشاط الأعمال.

الشكل سلسلة القيمة للبيانات الكبيرة

بعض خوارزميات تعلم الآلة المستخدمة في منصة البيانات الضخمة

خوارزميات التنبؤ بالسلاسل الزمنية (Time Series Forecasting Algorithms):

هناك العديد من الخوارزميات التي تستخدم للتنبؤ بالسلاسل الزمنية وتم التركيز في الدراسة على خوارزميتين أساسيتين:

1. **خوارزمية الانحدار الذاتي (Auto Regressive)**

نموذج الانحدار الذاتي (نموذج AR) هو في الأساس الطريقة المستخدمة لنمذجة سلوك مستقبلي أو حالي في سلسلة زمنية ، باستخدام بيانات من

السلوكيات السابقة في نفس السلاسل الزمنية. العملية هي في الأساس انحدار خطي للأداء المتغير في السلسلة الزمنية الحالية مقابل الأداء السابق لمتغير واحد في السلسلة. تشير السلسلة الزمنية في هذا السياق إلى تسلسل نقاط البيانات المدرجة في الرسم البياني وعادة ما يتم أخذها بالترتيب الزمني ، على سبيل المثال ، ارتفاع المد والجزر في المحيط الذي يتم التقاطه في نقطة زمنية محددة. لذلك ، في نموذج AR ، يستخدم المرء البيانات السابقة في مثل هذه السلسلة الزمنية للتنبؤ بنمذجة السلوك المتوقع إذا كان هناك ارتباط بين القيم المحددة أو نقاط البيانات والقيم التي تسبقها وتتبعها. الانحدار الخطي هو نموذج إحصائي يفترض أن جميع الظواهر الطبيعية مرتبطة خطيًا ، أي أنها تتبع خطًا. على الرسم البياني ، فإن المتغيرات المستقلة والتابعة سترسم خطًا مستقيمًا عند محاولة تحديد العلاقة. وبالمثل ، ينتهي نموذج AR بإعطاء علاقة رسومية خطية [10].

طريقة العمل

تعطى علاقة الانحدار الذاتي بالشكل:

$$Yt = \emptyset 1Yt-1 + \emptyset 2Yt-2 + \ldots. + \emptyset pYt-p + at$$

Yt: يمثل قيم الظاهرة المدروسة في الزمن t.

$\emptyset 1$ ، $\emptyset 2$ ، $\emptyset p$ هي معلمات الانموذج .

t. $Yt-1, \ldots\ldots, Yt-p$: هي قيم الظاهرة المرتدة زمنيا خلال الزمن

at: الخطأ العشوائي المستقل ويسمى التشويش الابيض

2. خوارزمية تعزيز التدرج الأقصى (XGBOOST Extremely Gradient Boosting)

خوارزمية XGBoost: وهي خوارزمية مطورة عن خوارزمية شجرة التعزيز التدرجي تستخدم نموذج أكثر تنظيماً للتحكم في مشكلة التعلم الزائد، الأمر الذي يمنحها أداءً أفضل .تم تطوير هذه الخوارزمية في عام2016 حيث فازت بالمركز الأول في مسابقة كاغل Kaggle لتحليل البيانات وتم نشر ورقة بحثية بطريقة عملها. في تعريف ويكيبيديا ، تعزيز التدرج هو تقنية التعلم الآلي المستخدمة في مشاكل الانحدار والتصنيف ، فهي تولد نموذجًا للتنبؤ في شكل مجموعة من نماذج التنبؤ الضعيفة (عادةً ما تكون أشجار القرار). أصبحت شائعة في الأيام الأخيرة وتهيمن على التعلم الآلي التطبيقي ومسابقات Kaggle للبيانات المنظمة بسبب قابليتها للتوسع.

XGBoost هو امتداد لأشجار القرار المعزز بالتدرج (GBM) وهو مصمم خصيصًا لتحسين السرعة والأداء.

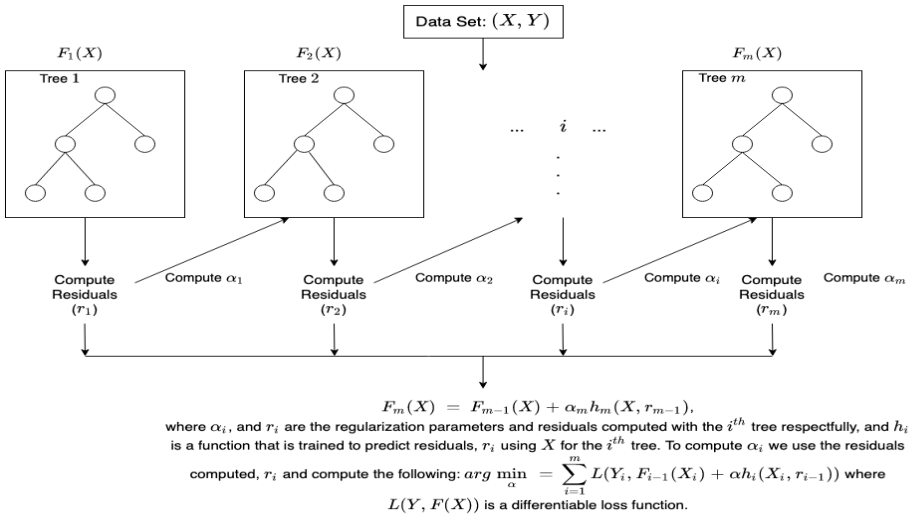
يستخدم XGBoost لمشاكل التعلم تحت الإشراف ، حيث نستخدم بيانات التدريب (مع ميزات متعددة) للتنبؤ بالمتغير المستهدف، حيث تحاول التنبؤ بدقة بالمتغير المستهدف من خلال الجمع بين تقديرات مجموعة من النماذج الأبسط والأضعف .[11]

طريقة العمل

يتكون التعلم الجماعي من مجموعة من المتنبئين (أشجار قرار) وهي نماذج متعددة لتوفير دقة تنبؤ أفضل. في تقنية التعزيز ، يستمر التدريب بشكل متكرر، بإضافة أشجار جديدة تتنبأ ببقايا أو أخطاء الأشجار السابقة التي يتم دمجها بعد ذلك مع الأشجار

السابقة لعمل التنبؤ النهائي. يطلق عليه تعزيز التدرج لأنه يستخدم خوارزمية النسب المتدرج لتقليل الخسارة عند إضافة نماذج جديدة.

تتم محاولة تصحيح الأخطاء التي ارتكبتها النماذج السابقة من خلال النماذج التالية عن طريق إضافة بعض الأوزان إلى النماذج.



الشكل يوضح مخطط عمل خوارزمية XGBOOST

معايير القياس والمفاضلة (Criteria for trade-offs between models):

هناك معايير متعددة يمكن عن طريقها تقييم الافضلية بين النماذج المتنبئ بها واعتمد الباحث على المعيار التالي:

جذر متوسط مربعات الأخطاء (Root Mean Squared Error RMSE)

إنه الجذر التربيعي لنسبة مجموع مربعات الانحرافات (الفوارق) بين الملاحظات والقيمة الحقيقية لعدد المشاهدات [12] .

يستخدم لقياس الانحراف بين القيمة المتنبئة والقيمة الحقيقية ومعادلتها من الشكل:

$$RMSE(X,h) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(h(x^{(i)}) - y^{(i)})^2}$$

بحيث أن :h(x) هي القيمة الحقيقة

y(i) هي القيمة المتنبئة

m عدد المشاهدات

4- النتائج ومناقشتها:

تم الحصول على بيانات من الانترنيت حيث استخدم الباحث ملفات CSV مختلفة:

بيانات الطقس (weather): تحتوي على مجموعة من السجلات recordsوالبالغ عددها 8760 تتضمن بيانات عن الطقس وسرعة الرياح ودرجات الحرارة والرطوبة خلال عام 2020.

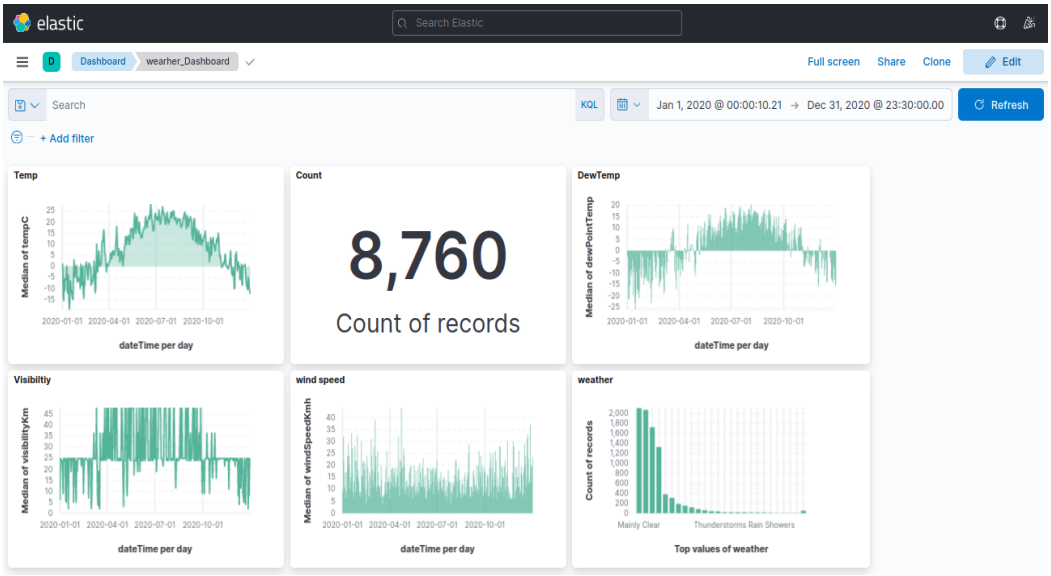
بيانات بيع منتجات (Products): تحتوي على مجموعة من السجلات recordsوالبالغ عددها 5000000 تتضمن معلومات عن مجموعة المنتجات ومناطق بيعها واجمالي الربح والتكاليف الخاصة بها خلال 15 عام.

تم إدخال هذه الملفات إلى elastic search باستخدام طريقة map reduce المضمنة مع Hadoop ، ومن ثم تم استخدام Kibana من أجل إنشاء Dashboard خاصة لكل ملف وتم فيها عرض النتائج الإحصائية مع تطبيق فلاتر معينة.

ومن ثم إجراء مقارنة بين نتيجة الخوارزمية Auto recursive لوحدها باستخدام متغير دخل واحد فقط ، وبين دمجها مع الخوارزمية xgboost باستخدام عدة متغيرات دخل للتنبؤ بالمتغير الهدف.

المثال الأول

نتائج تحليل Weather Data



يمثل الشكل جميع البيانات المتعلقة بالطقس خلال العام 2020

التغير الزمني باليوم

الشكل الأول يبين معدل تغير الحرارة مع الزمن

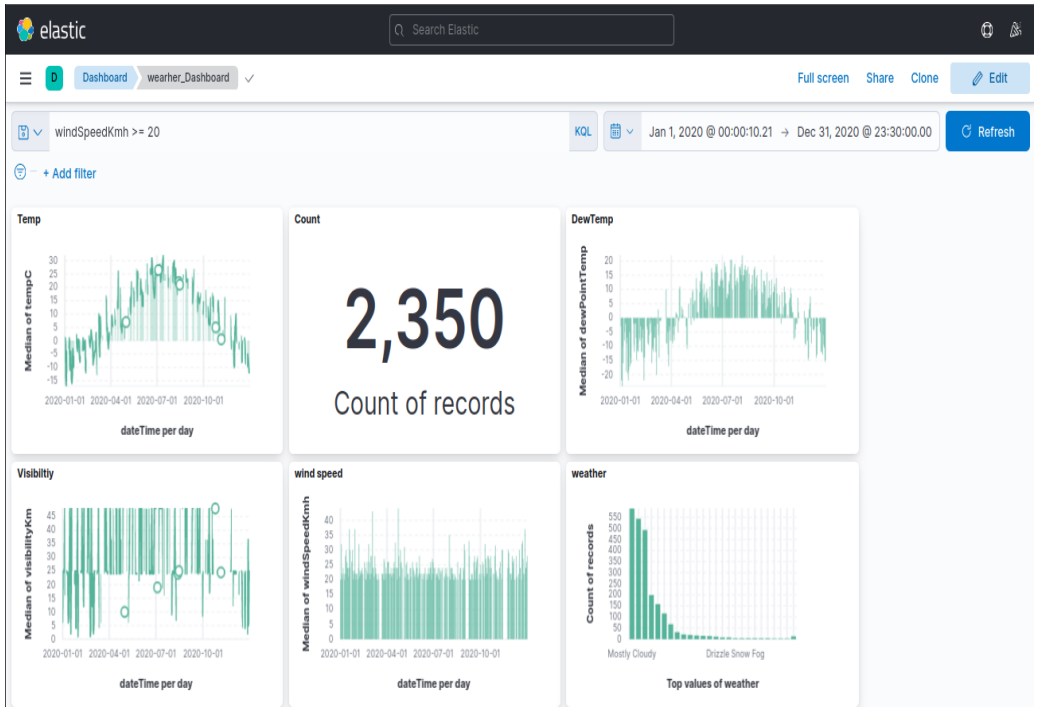
الشكل الثاني يمثل عدد السجلات

الشكل الرابع يبين معدل تغير الرؤية

الشكل الخامس يوضح تغير سرعة الرياح

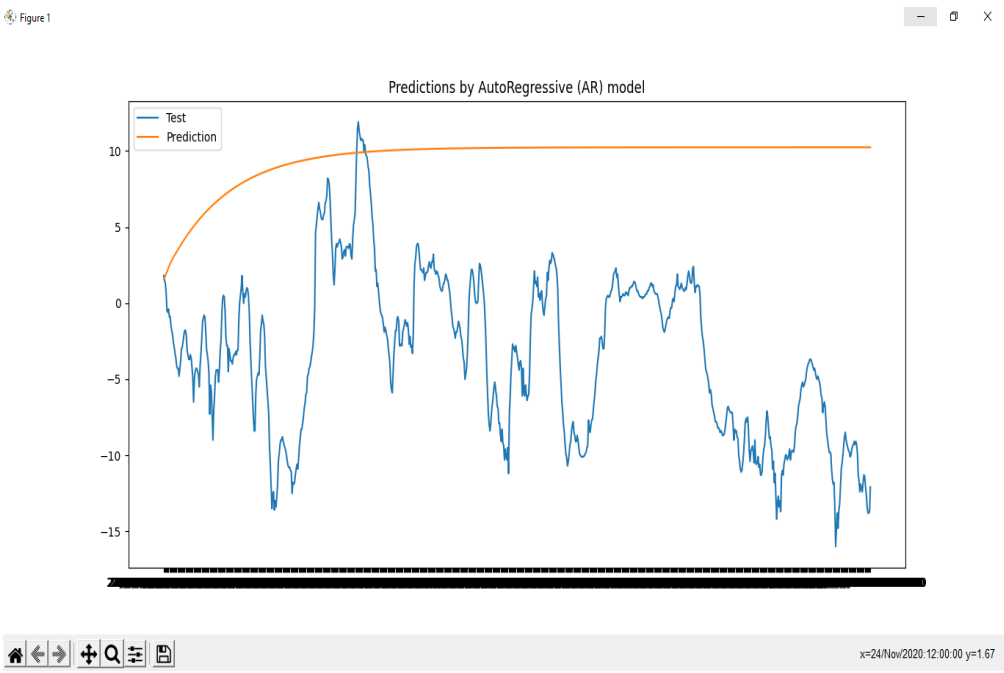
الشكل السادس يبين تصنيفات الطقس (ضباب – مثلج – ماطر – إلخ)

بعد تطبيق فلتر محدد (سرعة الرياح => 20) نلاحظ النتيجة التالية:



يمثل الشكل البيانات المتعلقة بالطقس بعد تطبيق فلتر محدد

تم استخدام خوارزمية auto recursive للتنبؤ بدرجة الحرارة بالاعتماد على متغير دخل واحد فقط واستعراض النتيجة.



الشكل يبين نتائج التنبؤ للمثال الأول باستخدام خوارزمية AutoRegressive

بحيث أن الخط البياني الأزرق يمثل القيم الفعلية للداتا والخط البياني الأخر يمثل القيم التنبئية. وقيم التنبؤ (العمود الأخير) ونسبة الخطأ

	Date	temp_c	dew_point_temp	rel_hum	...	visibility_km	press_kpa	weather	Predictions
7884	24/Nov/2020:12:00:00	1.8	-5.0	61	...	24.1	99.60	Snow Showers	1.593732
7885	24/Nov/2020:13:00:00	1.5	-7.4	52	...	24.1	99.64	Cloudy	1.668077
7886	24/Nov/2020:14:00:00	1.3	-6.3	57	...	24.1	99.65	Mostly Cloudy	1.782671
7887	24/Nov/2020:15:00:00	0.7	-6.4	59	...	2.4	99.70	Snow Pellets	1.932043
7888	24/Nov/2020:16:00:00	-0.5	-5.1	71	...	24.1	99.74	Mostly Cloudy	2.054483
...	...	...	...	...	...	...	...	...	...
8755	30/Dec/2020:19:00:00	-13.4	-16.5	77	...	25.0	101.47	Mainly Clear	10.241194
8756	30/Dec/2020:20:00:00	-13.8	-16.5	80	...	25.0	101.52	Clear	10.241196
8757	30/Dec/2020:21:00:00	-13.8	-16.5	80	...	25.0	101.50	Mainly Clear	10.241197
8758	30/Dec/2020:22:00:00	-13.7	-16.3	81	...	25.0	101.54	Mainly Clear	10.241198
8759	30/Dec/2020:23:00:00	-12.1	-15.1	78	...	25.0	101.52	Mostly Cloudy	10.241199

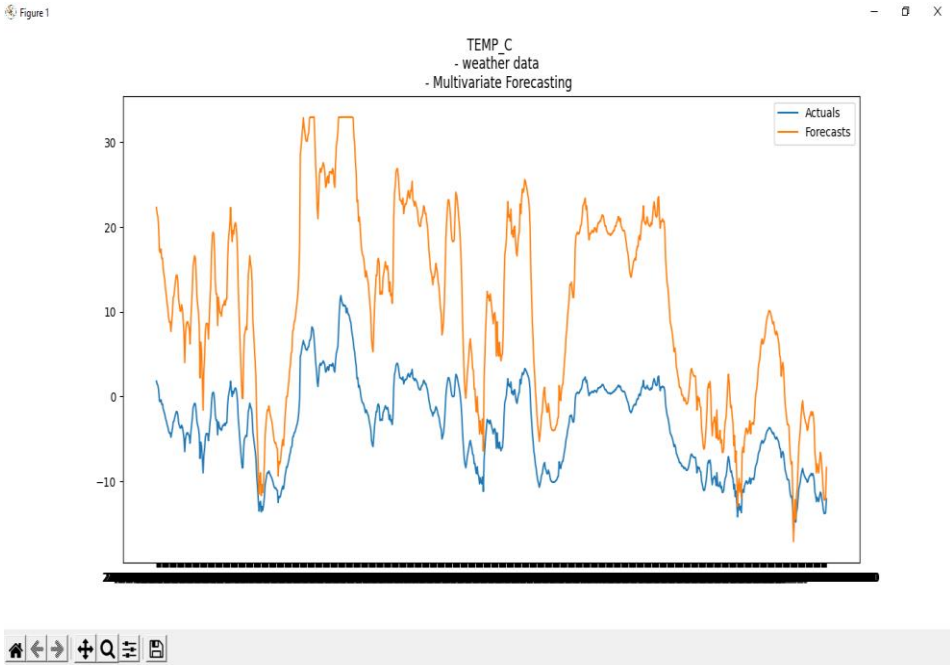
[876 rows x 9 columns]
AR - Root Mean Square Error (RMSE): 14.101

نلاحظ أن قيم التنبؤ أخذت منحى موجب تماما والمنحني أخذ منحى أسي

تم تعزيز الخوارزمية السابقة باستخدام خوارزمية XGBOOST وهذه الخوارزمية تأخذ عدة متغيرات دخل وتتنبأ بالخرج

في البداية تم استخدام خوارزمية AUTO REGRESSIVE للتنبؤ بمتغيرات الدخل (الرطوبة – سرعة الرياح – الضغط – الرؤية)

بعد تطبيق خوارزمية XGBOOST تم الحصول على المخطط التالي



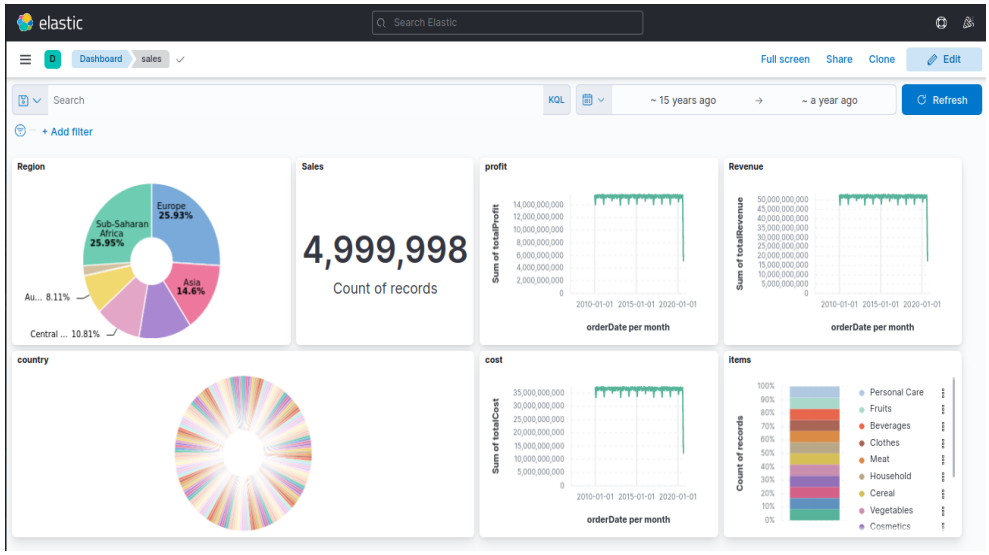
الشكل يبين نتائج التنبؤ للمثال الأول باستخدام خوارزمية Xgboost

ونسبة الخطأ للخوارزمية

```
XGBoost - Root Mean Square Error (RMSE): 15.296
```

المثال الثاني

نتائج تحليل Products Data



يمثل الشكل جميع البيانات المتعلقة ب Products على مدى خمسة عشر عام

يمثل الشكل الأول المنطقة

يمثل الشكل الثاني عدد التسجيلات

يمثل الشكل الثالث منحني تغير مجموع الأرباح مع الزمن (التغير الزمني شهريا)

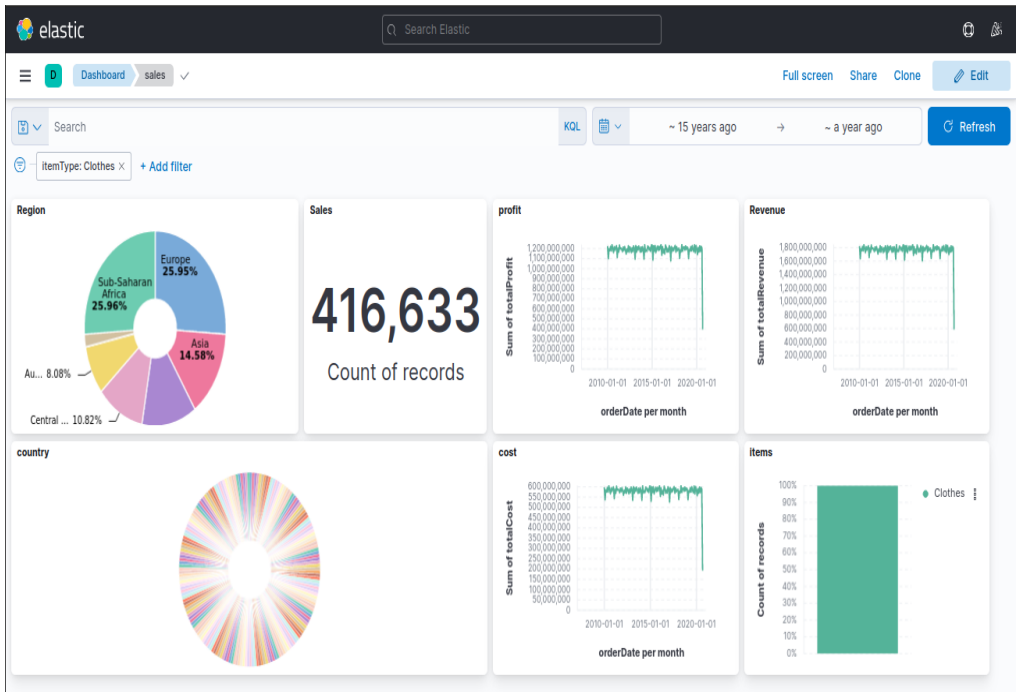
يمثل الشكل الرابع منحني تغير مجموع العائدات مع الزمن (العائدات = التكاليف + الأرباح)

يمثل الشكل الخامس البلدان

يمثل الشكل السادس منحني تغير مجموع التكاليف مع الزمن

يمثل الشكل السابع المنتجات

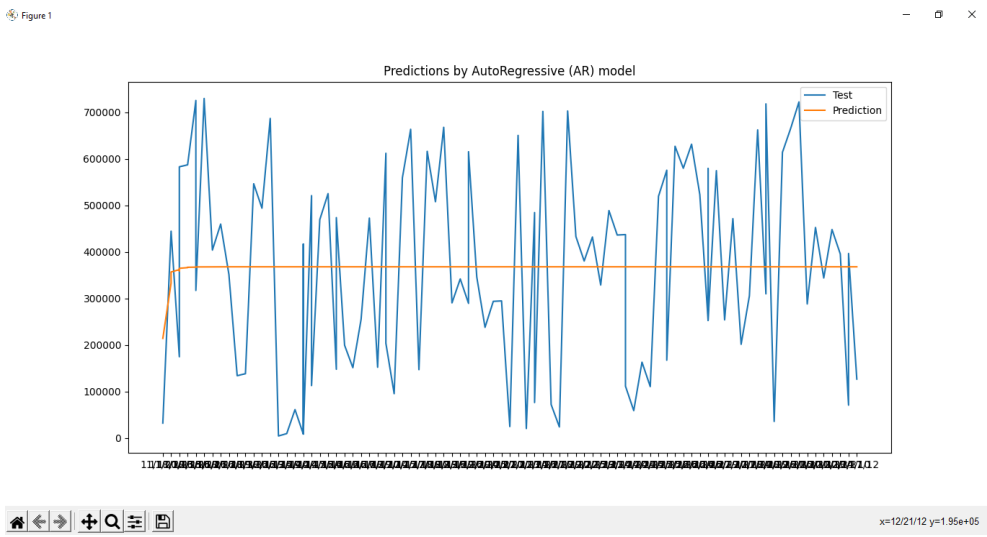
بعد تطبيق فلتر (المنتج = clothes)



يمثل الشكل البيانات المتعلقة ب Products بعد تطبيق فلتر محدد

تم فلترة البيانات بحيث يتم التنبؤ بالأرباح لمبيعات الملابس في القارة الأسيوية في بلد الصين

أولا باستخدام خوارزمية AUTO REGESSIVE



الشكل يبين نتائج التنبؤ للمثال الثاني باستخدام خوارزمية AutoRegressive

قيم التنبؤ ونسبة الخطأ:

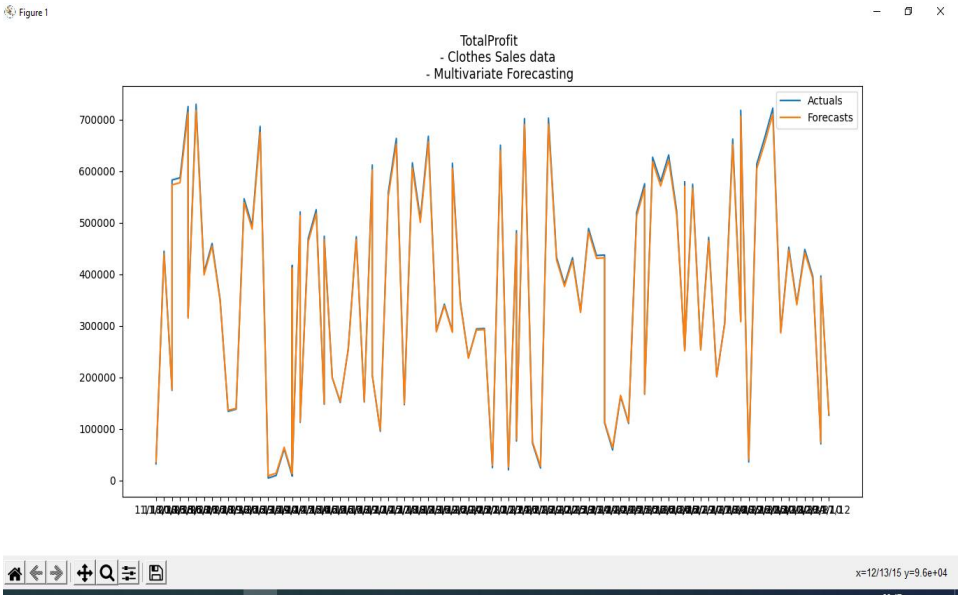
	Region	Country	ItemType	SalesChannel	OrderPriority	...	UnitCost	TotalRevenue	TotalCost	TotalProfit	Predictions
2030	Asia	China	Clothes	Online	C	...	35.84	47755.36	15662.08	32093.28	214229.286230
2031	Asia	China	Clothes	Online	H	...	35.84	661799.68	217047.04	444752.64	333244.738002
2032	Asia	China	Clothes	Online	H	...	35.84	661799.68	217047.04	444752.64	350616.688624
2033	Asia	China	Clothes	Online	H	...	35.84	661799.68	217047.04	444752.64	357032.463450
2034	Asia	China	Clothes	Online	H	...	35.84	259977.12	85263.36	174713.76	361994.850459
...	...	...	...	...	...	...	...	...	...	...	...
2251	Asia	China	Clothes	Online	L	...	35.84	590330.56	193607.68	396722.88	368134.871440
2252	Asia	China	Clothes	Online	L	...	35.84	105236.64	34513.92	70722.72	368134.871440
2253	Asia	China	Clothes	Online	L	...	35.84	590330.56	193607.68	396722.88	368134.871440
2254	Asia	China	Clothes	Offline	L	...	35.84	188398.72	61788.16	126610.56	368134.871440
2255	Asia	China	Clothes	Offline	L	...	35.84	188398.72	61788.16	126610.56	368134.871440

[226 rows x 15 columns]
AR - Root Mean Square Error (RMSE): 10.319

تم التحسين على الخوارزمية السابقة بإدخال متغيرين هما العائدات والتكاليف

حيث تم استخدام خوارزمية AUTO REGRESSIVE للتنبؤ بالمتغيرين أعلاه ثم تم إدخال النتيجة إلى خوارزمية XGBOOST

المخطط البياني بعد إدخال النتائج إلى خوارزمية XGBOOST



الشكل يبين نتائج التنبؤ للمثال الثاني باستخدام خوارزمية xgboost

ونسبة الخطأ

```
XGBoost - Root Mean Square Error (RMSE): 0.046
```

1. أثرت سرعة تحليل البيانات باستخدام منصة البيانات الضخمة بشكل مباشر على الحملة التسويقية كما في مثال (Products) وذلك من خلال إتاحة امكانية تحليل ملايين السجلات وتقديم النتائج في الزمن الحقيقي وبالتالي طرح العروض التسويقية في الزمن الحقيقي بالنسبة لسلوك الاستهلاك أو مكان وجود الزبائن.
2. تمكن منصة البيانات الضخمة من توليد سمات إضافية للمستخدمين وكان من أهم نتائجها توقع المنتجات الأكثر مبيعا من خلال تحليل سلوك المبيعات

واستخراج سمات تعبر عن هذه السلوكيات وهو ما يمثل ركيزة أساسية لمعظم الحملات التسويقية.

3. يمكن من خلال منصة البيانات الضخمة بناء نظام متكامل يحقق جميع مراحل دورة حياة البيانات من تحصيل البيانات ومعالجتها وتحليليها وتخزينها وحتى واجهات الويب التي تؤمن سهولة الاستخدام للمختصين.
4. ساهمت منصة البيانات الضخمة في تخفيض تكلفة استخدام العتاد الصلب بشكل كبير وذلك بسبب اعتمادها على استثمار الموارد بطريقة موزعة وبالتالي إمكانية استخدام مخدمات في نهاية عمرها مما يؤدي إلى تخفيض التكاليف عن التكاليف الفعلية.
5. يسهم استخدام بيانات التواصل الاجتماعي في توليد رؤى جديدة للشركات وتحسين جودة المنتجات مع تخفيض في التكاليف مما يسهم في زيادة المبيعات وزيادة القدرة التنافسية للشركات وبالتالي زيادة قيمة المنشأة.
6. إمكانية التنبؤ بالمشاكل قبل حدوثها وتحسين التنبؤ بالأرباح.
7. اهم معوقات انتشار تحليل البيانات الضخمة هو المخاطر المتعلقة بعنصر تحقيق الأمان نتيجة تعدد مصادر وضخامتها وبالتالي صعوبة تخزينها.
8. لم يتم قياس مدى تأثير تحليل البيانات بشكل فعلي على قدرات الديناميكية للشركة لأن البيانات التي تم تحليلها هي بيانات وهمية وليست حقيقية وبالتالي لم يستطع قياس مدى تأثيرها على قدرات الابتكار التزايدية والجزرية.
9. نلاحظ أن خوارزمية xgboost قد حققت أداء أفضل من خوارزمية autoregessive لاسيما أنها تأخذ عدة مدخلات وبالإضافة إلى نسبة الخطأ الأقل من نظيرتها.

5- الاستنتاجات والتوصيات:

1. زيادة البحث في مجال تحليل البيانات الضخمة في وسائل التواصل الاجتماعي وغيرها لمعرفة أهمية وتأثر هذه البيانات في الوسط المحيط والعمل على زيادة دقة النتائج.
2. إمكانية الاعتماد على النماذج المستخرجة من تحليل البيانات الضخمة لبناء نظام لتوليد العروض بشكل آلي وإعطاء العرض الأفضل بناء على التحليل.
3. يجب على الشركات التي تركز على الزبائن استخدام منصة تحليلات البيانات الضخمة لتحسين العروض المقدمة لزبائنها والتركيز على توقيت عمل زبائنها من أجل تحسين أرباحها.
4. إمكانية الاعتماد على منصة البيانات الضخمة في توقع الوقت الأفضل لنشر منتجات معينة بناء على نتائج تحليل سابقة وبالتالي زيادة الأرباح أيضا.
5. إمكانية الاستفادة من منصة البيانات الضخمة في تحديد فرص عمل جديدة وزيادة نمو المنظومات التي تركز على العملاء حيث يظهر اللجوء إلى خدمات جديدة حتى لو قدمت إلى جمهور صغير تمحورا حول العميل ينم عن عقلية مبتكرة.
6. وضع نتائج البحث أمام مختلف المنظمات التي تعتمد في عملها على تحليل البيانات من مصادر مختلفة والتي قد تتجه إلى تطبيق هذا النوع من التقنيات للاستفادة منها كالجهات الحكومية ومؤسسات الأعمال حيث يمكن أن تستفيد من هذه التحليلات في زيادة أرباحها وتعزيز وضعها التنافسي من خلال معرفة رغبات الزبائن وميولهم وبالتالي توفير منتجات وخدمات بناء على تلك الرغبات، وبالتالي تحقيق رضا العملاء بعيدا عن التخمين والمجازفة.
7. ضرورة قيام الشركات الكبيرة بتوفير الموارد اللازمة وعقد دورات تدريبة لتدريب المختصين على كيفية تحليل البيانات الضخمة كأحد أدوات التحول الرقمي وتوفير سبل أمان المعلومات.

6- المراجع:

1. Constantiou, I. D., & Kallinikos, J. (2015). New games, new rules: big data and the changing context of strategy. Journal of Information Technology, 30(1), 44-57.
2. 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.
3. McAfee, A., Brynjolfsson, E., & Davenport, T. H. (2012). Big data: the management revolution. Harvard Business Review, 90(10), 60-68.
4. Mikalef, P., & Pateli, A. (2017). Information technology-enabled dynamic capabilities and their indirect effect on competitive performance: Findings from PLS-SEM and fsQCA. Journal of Business Research, 70, 1-16.
5. Gupta, M., & George, J. F. (2016). Toward the development of a big data analytics capability. Information & Management, 53(8), 1049-1064.
6. Chen, C.P. & Zhang, C. Y. (2014) "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data", Information Sciences, pp. 275-314-347.

7. Zhou, J & Chen, P. & Chen, L. & Li, H.X & Zhao, W. (2013) "A collaborative fuzzy clustering algorithm in distributed network environments", IEEE Trans. Fuzzy Syst. PP. 99.
8. Yu Dong, Deng Li. (2011) "Deep learning and its applications to signal and information processing", IEEE Signal Process. Mag. 28 (1) pp. 145–154.
9. J.M. Cavanillas, E. Curry, and W. Wahlster. (2016) New Horizons for a Data-Driven Economy: A Roadmap for Usage and Exploitation of Big Data in Europe: Springer International Publishing, ISBN 9783319215686
10. https://thebusinessprofessor.com/en_US/research-analysis-decision-science/autoregression-definition
11. Chen Tianqi, Guestrin Carlos (2016) XGBoost: A Scalable Tree Boosting System, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 13-17, San Francisco, California, USA.
12. https://en.wikipedia.org/wiki/Root-mean-square_deviation