الجمهورية العربية السورية
وزارة التعليم العالي
المعهد العالي لإدارة الأعمال
قسم إدارة العمليات والمعلومات

استخدام تقانات تحليل المعطيات لدعم جودة النتائج الطبية

حالة عملية: إصابات فيروس COVID_19

The implementation of data analysis techniques in reinforcing the quality of medical results

(A case study based on Covid-19 cases)

إعداد الطالب
عبد الرحمن دياربكرلي

بإشراف

د. ياسر رحال **د. كادان الجمعة**

السنة الدراسية
الخامسة

العام الدراسي: 2020/2019

الشكر والإهداء

إلى أصحاب العقل الرشيد والركن الشديد والسند المتين والنصح الأمين من امنوا بنا

والدي و والدتي

إلى من كان سندا لي دون شر

إخوتي

إلى الداعمين و المتفانين معنا في طريق العلم

الدكتور كادان الجمعة

الدكتور ياسر رحال

المهندسة نظرة رحمة

إلى بحور العلم والمعرفة إلى العلماء العلامين اساتذتي:

د. وائل خنسة – د. طلال عبود – د. راضي خازم

إلى من كانوا سنداً في بناء هذا العمل المثمر من كانوا رفاق طريق العلم والمعرفة

أصدقائي

إلى شبابنا

الملخص :

تهدف هذه الدراسة إلى استخدام تقانات تحليل المعطيات وما يصاحبها من تقنية التنقيب في المعطيات , وذلك دعماً لجودة النتائج الطبية.

تأطيراً لتطبيق هذه الطرق والتقنيات والأدوات على مجموعة البيانات الخاصة بإصابات جائحة فيروس كورونا (**Covid_19**) لهدف تقديم نماذج رياضية من شأنها أن تكون مساعدة في دعم جودة تلك النتائج.

وقد تم الحصول على نوعين من مجموعات البيانات تلك الخاصة بمراكز:

(WHO and johns Hopkins university)

متضمنة (بيانات التسلسل الزمني للإصابات + الوفيات + حالات الشفاء بالإضافة إلى أعراض العدوى المصاحبة لفيروس كورونا) مجمعة على فترة زمنية تمتد **6** أشهر.

وبعد عملية جمع البيانات وتنظيفها , استخدم الباحث مجموعة من النماذج الرياضية و خوارزميات التنقيب في المعطيات

(linear regression – logistic growth model – time series model – Cluster analysis)

وذلك لتحليل تلك البيانات والوصول إلى نماذج تنبؤية مساعدة في دعم جودة النتائج الطبية.

ومن هذه النماذج استطاع الباحث أن يقدم تفسير حول كيفية الإستفادة منها في مجال الرعاية الصحية ليتمكن على أثرها من الوصول إلى عدة مراحل لدعم جودة النتائج الطبية وهي :

– تجميع بيانات أعراض الإصابات ضمن عناقيد ساعد في التنبؤ بالأنماط والمعلومات المخفية داخل البيانات الطبية .

– بناء نموذج لسلسلة زمنية يساعد في التنبؤ باتجاه الفيروس والتنبؤ بعدد الإصابات لفترات مستقبلية لمجموعة من أكثر البلدان تأثراً حول العالم.

– اقتراح نموذج لقياس نسبة نمو الفيروس وذلك للتنبؤ بالبلدان التي قد وصل عدد الإصابات فيها إلي الذروة

– اقتراح نموذج للتنبؤ بعدد الإصابات التي من الممكن وقوعها خلال تاريخ معين

ملاحظة الباحث:

الباحث ليس متخصص في الصحة أو أخصائيي في الأوبئة ، ولا ينبغي تفسير آراء هذا البحث على أنها نصيحة مهنية طبية بل هي مساعدة لتقدم نظرة من منظور رياضي أخر قد تكون مساعدة في عملية دعم جودة النتائج الطبية .

الفهرس

الفصل الأول

الإطار العام للبحث

(الإطار التمهيدي)

Data science and mining is changing the way we work

It's changing the way we use data

And it's changing the way organizations understand the world… **IBM**

المقدمة :

صاحب التقدم العلمي الحديث الذي شهدناه بشكل ملفت من بدايات القرن العشرين حتى يومنا هذا نمو كبير بالتقانات العلمية بشكل عام والتكنولوجية على وجه الخصوص.

فحُسنت الأدوات الموجودة بغية منحها القدرة للإجابة على الاسئلة المستجدة.

إلا أن الطيف الواسع للمشاكل المستجدة , وتعقيدها الناتج عن تعقيد وتطور عالمنا ومنظماتنا وتعقيد فهمنا لهذا العالم , جعل من تطوير الأدوات الموجودة أصلا عملاً غير كاف.

لذا فقد عمل العلماء (الرياضيين والمعلوماتيين خاصة منهم) إلى استحداث عدد كبير (وهو في زيادة دائما) من الأدوات والتقانات الجديدة المختبرة والمعقولة التي تزيد من قدرتنا على امتلاك المعرفة العلمية وتوسيع أفق تطبيقها إلى حدود غير مسبوقة في مؤسساتنا .

جلب لنا ذلك التقدم ثورة افتراضية جعلت من قدرتنا على التحكم بجوانب كثيرة من عالمنا قدرة عظيمة بلغت حدود ما كنا نعتبره خيالاً في يوما من الايام.

وكنتيجة لتلك الثورة الرقمية فقد ظهرت نظم المعلومات في السياق كإضافة حديثة للقائمة الطويلة من المناهج والممارسات والأدوات العلمية التي زودتنا بنظرة جديدة للكثير من مسائلنا التقليدية وساهمت في زيادة فهمنا لمنظومات عالمنا المعاصر وقدرتنا على تحليل ومعالجة وتصميم وحل مسائل هذه المنظومات بغية تمكننا من التحكم بها.

حيث أن نظم المعلومات , هي عبارة عن مجموعة متكاملة من المكونات لجمع البيانات وتخزينها وتجهيزها , وتوفير المعلومات والمعارف والمنتجات الرقمية.

حيث تعتمد الشركات ومراكز الأبحاث وغيرها من المنظومات على نظم المعلومات للقيام بعملياتها وإدارتها , والتفاعل مع زبائنها وموردیها , والتنافس في السوق وإثراء جوانب البحوث العلمية.

هذا في سياق تركيزنا على المعلومات في وقتنا الحاضر التي تمثل المورد الرقمي الأهم في المنظمة بشكل خاص والعالم بشكل عام , والذي تعتمد عليها في انجاز الوظائف الادارية (التخطيط، التنظيم، التوجيه، الرقابة) وغيرها الكثير من الاهميات المتفرعة فلن ننسى ما أحدثته المعلومات من تغييرات سياسية واقتصادية أقل ما يمكن وصفها بالتغيير الجذري في عالمنا كأحداث شركة Cambridge Analytica(2016).

فنجاح المنظمة يتوقف على كفاءة وفاعلية إدارتها في صنع القرارات , منطلقين من مفهوم جوهري اخر كون أن المعلومات هي الحجر الأساس الذي ترتكز عليه تلك القرارات في مختلف المستويات الإدارية ومختلف أقسام المنظمة كالتسويق – والموارد البشرية – والإنتاج وغيرها الكثير , مع ذكر الدور الذي قدمته نظم المعلومات في قطاعات اقتصادية وتعليمية وطبية كالجامعات والمشافي ومراكز البحوث الطبية في يومنا الحاضر التي تدعمها نظم المعلومات بشكل شبه كامل تقوم ببعض نشاطاتها عن طريق نظم معلومات محوسبة .

في سياق حديثنا عن المعلومات ونظم المعلومات والدور الذي قدمته , سنركز في بحثنا هذا عن إستثمار تقاناتها كتحليل المعطيات والتنقيب عن المعطيات ودورهما البارز في إثراء جودة النتائج الطبية , ليؤدي هذا إلى نشوء مجالات بحثية مستقلة كالتنقيب في المعطيات الطبية وتحليل المعطيات الطبية

Medical data Analysis – Medical data mining

حيث تسعى تقنيات تحليل / التنقيب في المعطيات الطبية الوصول إلى مجموعة البيانات الطبية واستخراج المعلومات المفيدة منها لتساعد على فهم أفضل وأعمق لهذه البيانات وتحسين عملية استخراج النتائج الطبية.

حيث ان التحليل والتنقيب في المعطيات الطبية يتبع ذات المنهج المتبع في الطرق التقليدية للتحليل والتنقيب في المعطيات .

ومنه يجب فهم البيئة التي سيتم التعامل معها في جمع البيانات ثم تنظيفها وترتيبها وانتقاء التقنيات التي يمكن تطبيقها وأخيراً تفسير النتائج والتحقق من مدى صحة التقنيات المطبقة .

كتعريف نظري هناك العديد من التعريفات ووجهات النظر ولكن الجميع يتفقون على أن تحليل المعطيات والتنقيب عن المعطيات هما مجموعتان فرعيتان متصلتان بنظم المعلومات كحالهم بالإتصال بذكاء الأعمال وعلم البيانات و تحليلات البيانات.

حيث يعد التنقيب عن البيانات عملية منهجية ومتسلسلة لتحديد واكتشاف الأنماط والمعلومات المخفية في مجموعة بيانات كبيرة. يُعرف أيضاً باسم اكتشاف المعرفة في قواعد البيانات.

أما على الصعيد الأوسع فإن تحليل المعطيات من ناحية أخرى ، هو مجموعة شاملة من استخراج البيانات التي تنطوي على استخراج البيانات وتنظيفها وتحويلها ونمذجة وتصورها بهدف الكشف عن معلومات مفيدة يمكن أن تساعد في استخلاص النتائج واتخاذ القرارات.

ينطلق هذا البحث من دراسة مفهوم الجودة في النتائج الطبية , ومن ثم ننتقل لتقديم لمحة عن تحليل البيانات و التنقيب في البيانات الطبية وخصوصيتها وأهم تطبيقاتها والطرق المستخدمة منها وصولاً إلى التعريف بالنماذج و الخوارزميات المستخدمة مثل:

Linear regression – logistic growth model – time series – clustering – Classification

وتطبيقاتها على البيانات الطبية .

حيث سواء شئنا أم أبينا فقد شهد العقد الأخير بأهتمام متزايد بإيجاد طرق لتحسين جودة النتائج الطبية , وقد ادى تطور نظم المعلومات إلى توافر بيانات كبيرة في المجال (الطبي)

فإستثمار هذه التقانات على صعيد الجوانب الطبية تشكل التوجه العالمي الأوحد للشركات والجامعات ومراكز البحث , لما له عدة ميزات مضافة على صعيد الرعاية الصحية .

فنرى العديد من الشركات الكبرى والجامعات ومراكز الأبحاث التي تستثمر اليوم

مواردها المعلوماتية في إثراء الجانب الطبية كمثال (Johns Hopkins University and Massachusetts Institute of Technology)

وقد تم إختبار إصابات (فيروس كورونا– coronavirus) كحالة عملية في هذا البحث لتوظيف تقانات تحليل البيانات عليها , بهدف تقديم مجموعة من المؤشرات إلى مراكز الأبحاث والمشافي التي قد تكون مقياس مساعد لهم في تحقيق نتائج أعمق على صعيد الرعاية الصحية.

وذلك لأسباب التالية:

- توفر البيانات والسهولة النسبية فالوصول إليها
- اهتمام المجتمع البحثي الطبي بهذا المجال في الوقت الحاضر وانفتاحه على أي مشاركة من شأنها أن تكون ورقة قوة في فهم سلوك الفيروس والقضاء عليه.
- ندرة الدراسات المطبقة في المعهد في مجال تحليل و تنقيب البيانات الطبية

وعليه، إن التوجه السائد في العالم حالياً وخاصة في القطاعات الطبية يتجه نحو استثمار أنظمة المعلومات وتقاناتها التي تساهم في تحقيق أهدافها , وهذا ما يقدمه البحث الحالي من خلال استخدام تقانات تحليل المعطيات لدعم جودة النتائج الطبية.

إشكالية البحث :

تسعى مراكز الأبحاث عالمياً في يومنا هذا لتكون السباقة في مجال البحث والتطوير على صعيد الجوانب الطبية على وجه العموم , وما يخص جائحة فيروس كورونا على وجهه الخصوص.

ففي سياق ملاحظة إنتشار جائحة (فيروس كورونا COVID_19-) في سوريا بشكل خاص والعالم بشكل عام , والتي تعني بظهور العديد من الأعراض المرضية المصاحبة للمصابين بهذا الفيروس .. التي أدت إلى وصول الإحصائيات في يومنا هذا إلى **16,628,029** مصاب و **655,845** متوفا و **10,217,285** معافى .

تسعى مراكز الأبحاث لتحليل البيانات الخاصة بالمصابين بهذا الفيروس , لتقدم نظرتها الإحصائية الرياضية على صعيد التنبؤ بالعديد من النتائج ضمن الإطار الرياضي أي أن نتائجها لا تعتبر ملزمة طبياً , أنما هي نظرة اخرى من جانب رياضي .

يتم تأطير ما يسعى إليه هذا البحث من بناء نموذج لسلسلة زمنية لدراسة كيفية إنتشار هذا الفيروس وما يصاحبه من زيادة أو نقصان في عدد الإصابات والتنبؤ بعدد الإصابات التي ستحدث في تاريخ مستقبلي لاحق .

كما يركز البحث على دراسة بيانات بعض البلدان وبناء نموذج يتنبأ بوصول بعض البلدان للذروة في عدد الإصابات دون بلدان أخرى .

لننتقل إلى دراسة الإصابات الناتجة عن هذا الفيروس وتجميع البيانات ضمن عناقيد

بناءً على الخصائص التي يمتلكونها لإكتشاف الأنماط والمعلومات المخفية فيها وتوزيع هذه العناقيد ضمن مجموعات تحمل كل منها معلومات مهمة لإستخراج البيانات , لتؤدي إلى توليد فرضيات حول المصابين بالفيروس وإرتباطهم بنوع الأعراض المصاحبة لهم.

هذا يستوجب تلخيص مشكلة البحث من خلال السؤال الرئيسي التالي :

كيف يمكن استخدام تقانات تحليل المعطيات بهدف القيام بعمليات التحليل والتنبؤ وبناء النماذج للبيانات الخاصة بالإصابات الناجمة عن فيروس كورونا لدعم جودة النتائج الطبية

حيث يشتق من هذا السؤال الرئيسي مجموعة من الأسئلة التالية :

i. هل يمكن تحسين جودة النتائج الطبية انطلاقاً من تجميع بيانات أعراض الإصابات ضمن عناقيد لإكتشاف والتنبؤ بالأنماط والمعلومات المخفية ؟

i i. هل يمكن تحسين جودة النتائج الطبية انطلاقاً من بناء نموذج لسلسلة زمنية للتنبؤ بإتجاه الفيروس والتنبؤ بعدد الإصابات لفترات مستقبلية ؟

i i i. هل يمكن اقتراح نموذج لقياس نسبة نمو الفيروس للتنبؤ بالبلدان التي قد وصل عدد الإصابات فيها إلي الذروة ؟

i v. هل يمكن إقتراح نموذج يتنبأ بعدد الإصابات / الوفيات / حالات الشفاء المتوقع حدوثها خلال يوم معين بناء على البيانات السابقة ؟

أهداف البحث :

يمكن تحديد أهداف البحث في الثلاث نقاط التالية :

- اقتراح مجموعة من المؤشرات والنماذج التي ستساعد المؤسسات الطبية ومراكز البحث على فهم السبب من إرتباط الأعراض الناتجة عن الفيروس بجنس المصاب وعمره وتوزعه الجغرافي .

- اقتراح نموذج للمساعدة في فهم كيفية سير إتجاه الفيروس في الانتشار للتنبؤ بعدد الإصابات لفترات مستقبلية

- اقتراح نموذج للمساعدة في إيضاح وضع البلدان من ناحية وصولها إلى حد الذروة في عدد الإصابات

- اقتراح نموذج للمساعدة في توقع عدد الإصابات – الوفيات – حالات الشفاء خلال يوم معين بناء على البيانات الزمنية السابقة .

"كل ما لا تستطيع قياسه، لن تستطيع إدارته"

Peter Drucker [1]

[1] Peter Ferdinand Drucher (1909 - 2005) was an Austrian-born American management consultant, educator, and author, whose writings contributed to the philosophical and practical foundations of the modern business corporation. He was also a leader in the development of management education, he invented the concept known as management by objectives and self-control, [and he has been described as "the founder of modern management".

أهمية البحث :

الأهمية النظرية / الأكاديمية :

- تتجلى الإضافة العلمية الأولى للمشروع كونه يقدمة دراسة حديثة في بلدنا سوريا في مجال تحليل معطيات جائحة فيروس كورونا وما يحمله من مخرجات قد تكون مفيدة للمجال المعرفي الطبي .

- قد يكون البحث إضافة قيمة على صعيد المراجع للمهتمين بالموضوع

- توفير منهجيات ونماذج عمل مساعدة لطلاب المعهد العالي كون الأخير يفتقر إلى هذا النوع من المشاريع ضمن مؤسسته التعليمية

الأهمية التطبيقية :

- اقتراح نموذج قابل للتعميم لسلسلة زمنية يمكن التنبؤ من خلالها بعدد الإصابات المتوقع حدوثها خلال فترات زمنية قادمة .

- اقتراح نموذج نمو لوجستي قابل للتعميم لوضع مقياس للبلدان التي قد وصلت إلى حد الذروة من الإصابات أم لا .

- اقتراح مقياس قابل للتعميم لمجموعة السمات المشتركة من أعراض الإصابات بين كل عنقود وبين الجنس والعمر والتوزع الجغرافي.

- يمكن لجهات خارجية كالمشافي ومراكز الأبحاث الإستفادة من نتائج البحث لمساعدة في تطوير نتائج مستقبلية أدق وأكثر تعمقاً .

حدود البحث :

تتلخص محددات البحث من خلال الحدود الزمانية والمكانية وفق الأتي :

محددات زمنية :

- تم إعداد البحث خلال الفترة الزمنية الممتدة ما بين 2020/7/9 إلى 2020/8/9

محددات مكانية / جغرافية :

- تم إنجاز البحث في الجمهورية العربية السورية , إلا أن المقاييس والنماذج المقدمة والمقترحة صممت ضمن إطار مجموعة بيانات عالمية لتتجاوز الحدود المكانية أي أن مخرجات البحث قد قدمت دراسات لبلدان خارج حدود الجمهورية العربية السورية (كوريا الشمالية – ايطاليا– روسيا) .

محددات موضوعية :

- تم إعتماد البحث على مجموعة بيانات :
(WHO and johns Hopkins university) وذلك لعدم توفر مصادر أخرى بنفس مستوى دقة البيانات .

معوقات البحث :

يمكن تلخيص المعوقات التي واجهت الباحث على الشكل التالي :

1- صعوبة الحصول على دراسات سابقة مشابهة لمنهجية البحث المستخدمة.

كون المشاريع المشابهة غالبا ما تتسم بالحماية الفكرية وبالتالي المعلومات التي يتم مشاركتها حول مجالها تبقى محدودة بشكل كبير.

2- غياب القوانين والبنية التحتية الداعمة التقنية لدى المؤسسات العامة والخاصة فيما يخص تواجد انظمة معلومات للقيام بتطبيق الأدوات التي يمتلكها الباحث عليها.

3- ضعف الوعي حول أهمية انظمة المعلومات واستخداماتها بمختلف القطاعات الاقتصادية والخدمية.

4- غياب الدعم التقني المالي والفكري لنقل فكرة هذا المشروع من ضمن إطارها كعمل فردي إلى التحول كعمل فريق بحثي متكامل كما هو الحال في جامعة MIT التي خصصت جناح من حرم الجامعة للقيام بتجارب تحليل البيانات على فكرة مشابهة لهذا البحث (إصابات فيروس COVID_19).

الدراسات السابقة :

لقد اعتمد البحث بشكل أساسي على الدراستين التاليتين:

- **الدراسة الأولى (2014, Eapen م):**

أطروحة أعدت لنيل درجة الماجستير في هندسة تصميم النظم من قبل الباحث

(Arun George Eapen) خلال عام 2014 في كندا

بعنوان: "Application of Data mining in Medical Applications"

"تطبيق التنقيب في البيانات في التطبيقات الطبية"

حيث إن هذه الدراسة تهدف في جانبها النظري إلى استخدام أدوات التنقيب في البيانات ضمن تطبيقات الرعاية الصحية لتطوير أداة يمكن أن تساعد في صناعة قرارات دقيقة في الوقت المناسب.

أما بالنسبة الى جانبها العملي فقد تجلت الأهمية من خلال التطبيق العملي لخوارزميات التصنيف الخاصة بالتنقيب في المعطيات على قواعد المعطيات الطبية

و تزويد خوارزميات التعلم بمجموعة تدريب جيدة يتم من خلالها استخراج الأنماط للمساعدة في تصنيف مجموعة بيانات الاختبار.

(إن دراسة الحالة كانت تتكون من 455 سمة وأكثر من 4000 من الحالات)

ثم خلصت الدراسة إلى مجموعة من النتائج , أهمها :

إن خوارزميات ذكاء الآلة تتحسن مع أدوات استخراج البيانات , وإن بيانات مجال الرعاية الصحية يعد حالة عملية جيدة لاستخراج البيانات.

في هذه الأطروحة , استخدم الباحث خوارزمية ZeroR , وفسر ذلك بأن الخوارزمية

في بعض الأحيان تفوقت على بعض الخوارزميات الأخرى لاستخراج البيانات , وذلك يعود بسبب أن ZeroR تعتمد على الهدف وتتجاهل جميع المتنبئات وتعتبر الخوارزمية مفيدة لتحديد أداء خط الأساس كمعيار لطرق التصنيف الأخرى.

أما دقة المصنف فقد حصل على 75.75٪ من التجربة الأولى

وهو ما يعني بأن 75.75٪ من بيانات الاختبار ، يمثلون أغلبية مجموعة التدريب.

كما عمل الباحث على استخدام خوارزمية Naïve Bayes حيث تعد هذه خوارزمية شائعة الاستخدام لإنتاج نتائج مصنفة بسرعة عالية جدًا.

يأتي التنبؤ الدقيق مع خوارزمية Naïve Bayes عندما يكون جميع المتغيرات المستقلة مستقلة إحصائيا عن بعضها البعض.

بالمقابل كانت تجارب شجرة القرار التي تم إجراؤها هي الأكثر فائدة و الغنية بالمعلومات.

تم الاستفادة من الدراسة السابقة بما يلي :

قدمت الدراسة السابقة إلى البحث القائم نظرة تطبيقية معمقة عن آلية إستخدام أدوات تنقيب المعطيات كبرنامج Weka وقد أظهر تصور واضح لعملية تنقيب البيانات عبر البرنامج.

كما تحدث الجانب النظري من الأطروحة عن الأوبئة وكيفية تحليل بياناتها , كون أن البحث المقدم من طرف الباحث يدور أيضًا في فضاء الأوبئة (فيروس كورونا).

- **الدراسة الثانية (رحمة , ٢٠١٧ م):**

أطروحة أعدت لنيل درجة الماجستير في المعلوماتية اختصاص نظم المعلومات ودعم القرار من قبل الباحثة:

نظرة أنطون رحمة خلال عام ٢٠١٧ في الجمهورية العربية السورية

بعنوان : "استخدام تقنيات التنقيب في البيانات لدعم جودة التعليم "

حالة عملية : المعهد العالي لإدارة الأعمال HIBA

حيث يندرج هذا البحث في سياق الأبحاث التي تعمل على ضبط جودة مدخلات العملية التعليمية وتحسين آليتها لضمان مخرجات هذه العملية ممثلة بمستوى الطالب الخريج .

واعتمد الباحث في اختيار مكان تطبيق بحثه في المعهد العالي لإدارة الأعمال HIBA

ليكون مجالاً للتطبيق , حيث قدم البحث مجموعة من المؤشرات التي تساعد إدارة المعهد على تشخيص مدى فاعلية بعض عناصر مدخلات العملية التعليمية وتوجيهها نحو وضع أسس لتطوير هذه العناصر بما يضمن زيادة الفاعلية , وتم ذلك عبر توظيف طرق التنقيب في البيانات التعليمية ومفهوم المنطق الغائم والمنطق الرياضي لتقديم نماذج مختلفة يتم تطبيقها حسب نمط المسألة التي تتم معالجتها

وتناول البحث ٣ مسائل أساسية :

١ – دراسة فاعلية قواعد القبول في ضمان مخرجات العملية التعليمية حالة المعهد العالي لإدارة الأعمال

٢– دراسة فاعلية الخطط الدراسية حالة المعهد العالي لإدارة الأعمال

٣- اقتراح نموذج حصر الطلاب المتوقع تخرجهم بمستوى حرج من مراحل مبكرة من مسارهم الدراسي بهدف التقويم والتوجيه حالة المعهد العالي لإدارة الأعمال.

ثم خلصت الدراسة الى مجموعة من النتائج أهمها :

وصل الباحث لنموذج يساعد على توقع مستوى الطلاب بشكل تقريبي لأنه من الصعب جداً توقع نتيجة الطلاب بدقة لعدة أسباب منها المرحلة المبكرة جداً التي يتم التوقع عندها وذلك بهدف التقويم من وقت مبكر , ولانه يوجد العديد من العوامل التي تدخل في مسيرة الطلاب منها ما له علاقة بإدارة العملية التعليمية ضمن المؤسسة ومنها له علاقة بميول الطلاب وحالته النفسية وعوامل أخرى عديدة.

لذلك من الأفضل إعادة تطبيق النموذج بالتوازي مع تقدم الطلاب أكاديمياً بمعنى إعادة تطبيق النموذج في نهاية السنة الدراسية الثانية مع إدخال مؤشرات جديدة للوصول إلى توقعات أكثر دقة واكتشاف العوامل التي من الممكن أن تكون مؤثرة في مستوى الطلاب عند التخرج ومحاولة تصحيحها.

تم الاستفادة من الدراسة السابقة بما يلي :

قدمت الدراسة السابقة إلى البحث القائم معرفة نظرية معمقة حول مفهوم الجودة حيث يعد مفهوم الجودة هو الركيزة الأساسية المراد الوصول لها عن طريق البحث القائم , كما تم الإستعانة في بحث (رحمة , ٢٠١٧ م): كنموذج أولي لكيفية كتابة البحوث العلمية , من هيكل عام للبحث لكيفية توزيع الفصول وصولاً إلى النتائج والتوصيات.

الفصل الثاني

الإطار النظري

(المبحث الأول: التقانات – المنهجيات – التقنيات – الخوارزميات – الأدوات)

"The most beautiful experience we can have is the mysterious"

Albert Einstein

المحتوى:

يتضمن الإطار النظري المفاهيم الأساسية التي تؤطر عمل الإطار العملي للبحث وهي:

- علم البيانات – **Data Science**
- تحليلات البيانات – **Data Analytics**
- تقانات تحليل المعطيات – **Data Analysis**
- تقنية التنقيب في المعطيات – **Data Mining**
- تقنيات علوم البيانات – **Data Science Techniques**
- خوارزميات علوم البيانات – **Data Science Algorithms**
- تقنيات تحليل البيانات – **Data Analysis Techniques**
- منهجيات / طرق التنقيب في المعطيات – **Data mining methods**
- خوارزميات التنقيب في المعطيات – **Data mining algorithms**

علم البيانات - Data science

تعريف :

علم البيانات (المعروف بالنموذج الرابع للعلوم)[٢]

بداية بقفزة زمنية إلى الوراء نحيط أن كلمة "علوم البيانات" كانت موجودة منذ حوالي الستينيات , إلا أنه في ذلك الوقت تم استخدامها كبديل لـ "علوم الكمبيوتر", الذي يحملان اليوم معنيين مختلفين.

في عام 2008 ، أصبح D.J Patil و Jeff Hammerbacher أول من أطلقوا على أنفسهم اسم "علماء البيانات" لوصف دورهم في LinkedIn و Facebook على التوالي.

في عام 2012 , بمراجعة لمقالة Harvard Business Review تم الإشارة إلى علم البيانات على أنه "الوظيفة الأكثر جاذبية في القرن الواحد والعشرين".

علم البيانات أشبه بعملية متواصلة ليست حدثاً منفصلاً , وهي عملية إستخراج البيانات لاستيعاب مختلف الأمور , ولفهم العالم على نحو أفضل.

حيث يؤطر علم البيانات في أنه فن الكشف عن الرؤى والاتجاهات التي تتوارى خلف البيانات , فكما تعني علوم الأحياء بدراسة الأحياء والعلوم الفيزيائية تدرس العلاقات الفيزيائية في العالم , فإن علم البيانات هو معني بدراسة البيانات.

برز ظهور علم البيانات وأهميته وذلك نظراً لتراكم وانفجار كميات هائلة من البيانات في زمننا الحاضر , كما كنا في السابق نعاني من نقص البيانات .. اليوم لدينا ترف في البيانات , ولم نكن نمتلك خوارزميات .. اليوم لدينا خوارزميات في متناولنا , كما

(2) Data Science has been referred to as the fourth paradigm of Science. (The other three being Theoretical, Empirical and Computational).

كانت البرامج غالية السعر .. فاليوم أصبحت بشكل شبه مجاني , والأهم من ذلك أن التخزين كان مشكلة .. اليوم لدينا مساحات تخزين غير محدودة .

لننتقل بعد ذلك إلى أهميتها للمؤسسات , حيث يساعد علم البيانات المؤسسات على فهم بيئتها وتحليل المشاكل التي تواجها للإضافة إلى قائمة المعرفة المجدولة التي تتمتع بها المؤسسة , حيث يستخدم علماء البيانات أدوات كال Data visualization

وذلك لمساعدة ال Stakeholder في فهم النتائج المقدمة من قبل علماء البيانات.

في نظرة تاريخية سريعة إلى الماضي لعلم البيانات وماهية الأسس التي بني عليها نلاحظ أن معظم مكونات علم البيانات من احتمالات – احصاء – جبر – الجبر الخطي – البرمجة – قواعد المعطيات – كلها كانت موجودة منذ عقود و لكن اليوم لدينا القدرات الحسابية لتطبيقها ودمجها والخروج بتقنيات جديدة و خوارزميات تعلم ذكي.

وكما ذكرنا آنفاً علوم البيانات وقدرتها على تقديم حلول وتحاليل وقيمة مضافة إلى ال Stakeholder في مؤسسات الأعمال , فعلينا عدم تغييب كيف أن علم البيانات كان وما يزال يستخدم في مجالات الرعاية الصحية.

حيث إن إستخدام تقنيات علوم البيانات وذلك لتحليل مجموعة البيانات الكبيرة المتاحة له تأثير كبير على الحياة البشر عن طريق :

- توفير معلومات موجهة لمساعدة المختصين في الرعاية الصحية لتقديم أفضل علاج , كمثال الدراسة القائمة حالياً من قبل الباحث حول فيروس كورونا , الأمر الذي دعا العديد من مراكز الأبحاث عالمياً للطلب من علماء البيانات النظر والتدقيق وتحليل هذه البيانات للوصول إلى معرفة مرجوة قد تكون مفيدة للمجتمع الدولي في مجال الرعاية الصحية .

ولن ننسى ما يستخدمه علماء البيانات من تقانات أخرى مرتبطة إرتباط وثيق بعلم البيانات للمساعدة في فهم أدق لبيانات الرعاية الصحية وغيرها كحال :

Data Modeling –Data mining –Statistics- Data Analysis – ML -DL)

تقاطعات علم البيانات :

في الأساس علم البيانات هو تقاطع ثلاث مجالات وهي:

1. الإحصائيات: يلعب هذا دورًا حيويًا لأن الرياضيات هي جوهر علم البيانات.

2. تحليل البيانات: يتم استيرادها أيضًا نظرًا لأن البيانات تحتاج إلى تحليل ورسم لتحديد تعقيداتها.

3. التعلم الآلي: يتألف من خوارزميات مختلفة تتضمن الإحصائيات.

دورة حياة علم البيانات :

تتمحور دورة حياة علوم البيانات حول استخدام التعلم الآلي والأساليب التحليلية الأخرى لإنتاج رؤى وتنبؤات من البيانات من أجل تحقيق هدف العمل.

تتضمن العملية بأكملها عدة خطوات مثل تنظيف البيانات , والتحضير , والنمذجة , وتقييم النموذج , وما إلى ذلك.

فهي عملية طويلة وقد تستغرق عدة أشهر حتى تكتمل.

متابعة الهيكل العام CRISP-DM لسير العملية بشكل قياسي الذي يتضمن:

1. فهم الأعمال – Business Understanding

تدور دورة حياة علم البيانات بأكملها حول هدف العمل لذا من المهم فهم هدف العمل بوضوح لأن ذلك سيكون الهدف النهائي من التحليل

2. فهم البيانات – Data Understanding

وهي عملية جماعية ضمن فريق عمل متكامل لجمع كل البيانات المتاحة ووصفها , كما تستكشف البيانات باستخدام الرسوم البيانية لاستخرج أي معلومة منها

3. إعداد البيانات – Data Preparation

تتضمن هذه الخطوة اختيار البيانات ذات الصلة , ودمج البيانات عن طريق دمج مجموعات البيانات ، وتنظيفها ، ومعالجة القيم المفقودة ومعالجة البيانات الخاطئة والتحقق أيض ًا من القيم المتطرفة والتعامل معها .

4. تحليل البيانات الاستكشافية – Exploratory Data Analysis

تتضمن هذه الخطوة الحصول على فكرة عن الحل والعوامل المؤثرة عليه ، قبل بناء النموذج الفعلي.

5. نمذجة البيانات – Data Modelling

نمذجة البيانات هي جوهر تحليل البيانات , يأخذ النموذج البيانات المعدة كمدخل ويوفر المخرج المطلوب.

تتضمن اختيار نوع النموذج المناسب , سواء كانت المشكلة مشكلة تصنيف أو مشكلة انحدار أو مشكلة عناقيد.

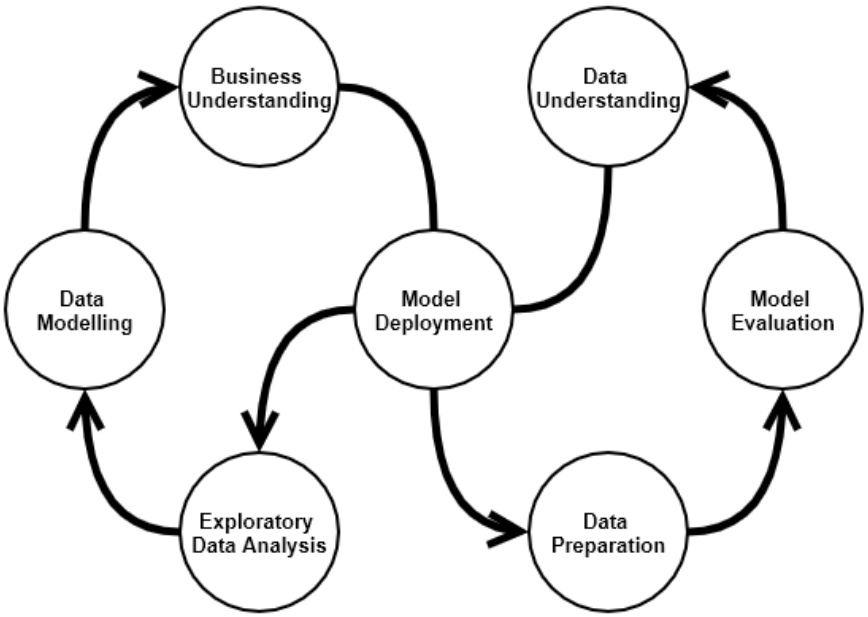
ثم أختيار الخوارزمية المطلوبة

6. تقييم النموذج – Model Evaluation

تقييم النموذج للتحقق مما إذا كان جاهز ًا للنشر عن طريق اختبار النموذج بناء ً على بيانات غير مرئية ، ويتم تقييمه بناء ً على مجموعة مدروسة بعناية من مقاييس التقييم. لنتأكد أن النموذج يتوافق مع الواقع.

7. نشر النموذج

المرحلة الأخيرة .


Business Understanding
Data Understanding
Data Modelling
Model Deployment
Model Evaluation
Exploratory Data Analysis
Data Preparation

تحليلات البيانات – Data Analytics

كما تناولنا في الفقرة السابقة من المبحث الأول عن علوم البيانات نستعرض هنا ماهية تحليلات البيانات .

حيث أنه كما أسلفنا بأن علم البيانات هو دراسة من أين تأتي المعلومات ، وما تمثله وكيف يمكن تحويلها إلى مورد قيم.

ليأتي دور تحليلات البيانات , يشبه علم البيانات , ولكن بطريقة أكثر تركيزاً .

الغرض من تحليلات البيانات هو توليد رؤى من البيانات من خلال ربط الأنماط والاتجاهات بالأهداف التنظيمية. تستخدم Data Analytics تعبيرات استعلام أساسية مثل SQL لتقطيع البيانات وتقسيمها.

إن البيانات التي تم إنشاؤها من مصادر مختلفة مثل السجلات المالية والملفات النصية وأشكال الوسائط المتعددة وأجهزة الاستشعار والأدوات هي بيانات كبيرة Big Data. أدوات ذكاء الأعمال البسيطة غير قادرة على معالجة هذا الحجم الضخم ومجموعة متنوعة من البيانات.

هذا هو السبب في أننا نحتاج إلى أدوات وخوارزميات تحليلية أكثر تعقيداً وتقدماً للمعالجة والتحليل ورسم رؤى ذات معنى للخروج منها.

في مقارنة بسيطة .. ينظر علماء البيانات بشكل أساسي إلى مجموعات واسعة من البيانات حيث يمكن أو لا يمكن إجراء الاتصال بسهولة بينما تنظر Data Analytics في مجموعة معينة من البيانات للتواصل بشكل أكبر.

يستخدم مجال علوم البيانات الرياضيات والإحصاء وتخصصات علوم الكمبيوتر ، ويتضمن تقنيات مثل التعلم الآلي وتحليل الكتلة واستخراج البيانات والتصور بينما تعمل تحليلات البيانات على لغة الاستعلام الهيكلية مثل SQL / Hive لدفع المخرجات النهائية.

تعمل تحليلات البيانات على تحليل الأسئلة التي يطرحها العمل بينما سيقوم عالم البيانات بصياغة الأسئلة التي من المرجح أن تفيد حلولها الأعمال

يتم تصنيف عملية تحليلات البيانات ذاتيًا إلى ثلاثة أنواع بناءً على الغرض من . تحليل البيانات كـ

تحليلات وصفية – التحليلات التنبؤية – التحليلات الإلزامية

مراحل معالجة البيانات:

1 – استخراج البيانات

2–تنظيف البيانات وتحويلها

3–تحليلها

4–تصور البيانات

تعريف :

في عصرنا الغني بالبيانات المسمى عصر إنفجار البيانات , يعد فهم كيفية تحليل واستخراج المعنى الحقيقي للأفكار والمعرفة من تلك البيانات هو المفتاح الأساسي للنجاح على صعيد المجالات الطبية مجالات الأعمال وغيرها الكثير.

فيُعرّف تحليل البيانات على أنه عملية تنظيف البيانات وتحويلها ونمذجتها بهدف اكتشاف معلومات مفيدة لعملية صناعة القرارات المستنيرة .

ويمكننا تأطير الغرض من تحليل البيانات في قدرته على استخراج معلومات مفيدة من البيانات واتخاذ القرار بناءً على تحليلات تلك البيانات .

استخدام تحليل البيانات :

يُستخدم تحليل البيانات في اّفاق الأعمال والاّفاق الطبية وغيرها الكثير لمساعدة المؤسسات على اتخاذ قرارات عمل أفضل.

سواء كان ذلك يتعلق بأبحاث السوق ، أو أبحاث طبية ، أو تحديد الموقع , أو مراجعات العملاء , أو تحليل المشاعر , أو أي مشكلة أخرى تتعلق بالبيانات , فإن تحليل البيانات سيوفر رؤى تحتاجها المؤسسات من أجل اتخاذ الخيارات الصحيحة.

كما يُعد تحليل البيانات أمرًا مهمًا للشركات اليوم , لأن الاختيارات التي تعتمد على البيانات هي الطريقة الوحيدة لتكون واثقًا حقًا في قرارات العمل. قد يتم إنشاء بعض الشركات الناجحة على حدس ، ولكن معظم خيارات الأعمال الناجحة تقريبًا تعتمد على البيانات.

أنواع تحليل البيانات :

1 – التحليل الوصفي / Descriptive Analysis:

وهو جزء من (Statistical Analysis) يبحث تحليل البيانات الوصفية في البيانات السابقة ويخبر ما حدث , حيث يقدم فكرة عن توزيع البيانات مؤشرات الأداء والإيرادات والعملاء المحتملين والمزيد كما يساعدك على اكتشاف القيم المتطرفة والأخطاء المطبعية.

2– التحليل التشخيصي / Diagnostic Analysis:

يهدف تحليل البيانات التشخيصية إلى تحديد سبب حدوث شيء ما. بمجرد أن يظهر التحليل الوصفي الخاص بنا حدوث شيء سلبي أو إيجابي ، يمكننا إجراء التحليل التشخيصي لمعرفة السبب.

مثال: قد يرى نشاط تجاري زيادة العملاء المحتملين في شهر تشرين الأول (أكتوبر) واستخدام التحليل التشخيصي لتحديد الجهود التسويقية التي ساهمت بشكل أكبر.

3– التحليل التنبؤي / Predictive Analysis:

يتنبأ تحليل البيانات التنبؤية بما يمكن أن يحدث في المستقبل.

حيث تُستمد الاتجاهات من البيانات السابقة التي تُستخدم بعد ذلك لتشكيل تنبؤات حول المستقبل.

على سبيل المثال ، للتنبؤ بعدد الإصابات بفيروس (X) للعام المقبل ، سيتم تحليل بيانات السنوات السابقة.

4– التحليل الإلزامي / Prescriptive Analysis:

يجمع تحليل البيانات الإلزامية المعلومات الموجودة من الأنواع الثلاثة السابقة لتحليل البيانات ويشكل خطة عمل للمنظمة لمواجهة المشكلة أو القرار.

هذا هو المكان الذي يتم فيه اتخاذ الخيارات القائمة على البيانات.

5- التحليل الاستدلالي / Inferential Analysis:

تحاول الوصول إلى استنتاجات تتجاوز البيانات الفورية وحدها , اي العثور على استنتاجات مختلفة من نفس البيانات بتحديد عينات مختلفة .

فعلى سبيل المثال نستخدم إحصائيات استنتاجية لمحاولة الاستدلال من عينة البيانات على ما قد يعتقده السكان.

أو لإصدار أحكام باحتمالية أن الاختلاف الملحوظ بين المجموعات يمكن الاعتماد عليه أو ربما حدث صدفة في هذه الدراسة , وبالتالي يستخدم لعمل استنتاجات من بياناتنا لظروف أكثر عمومية

عملية تحليل البيانات :

تتلخص بكونها ليست سوى عملية جمع المعلومات باستخدام تطبيق أو أداة مناسبة تسمح للباحث باستكشاف البيانات والعثور على نمط فيها.

بناءً على تلك المعلومات والبيانات , يمكننا اتخاذ القرارات ، أو يمكننا الحصول على استنتاجات نهائية.

يتكون تحليل البيانات من المراحل التالية:

1	Data Requirement Gathering	جمع متطلبات البيانات
2	Data Collection	جمع البيانات
3	Data Cleaning	تنظيف البيانات
4	Data Analysis	تحليل البيانات
5	Data Interpretation	تفسير البيانات
6	Data Visualization	عرض مرئي للمعلومات

1– جمع متطلبات البيانات :

ترتكز على الإستفسار ماذا نريد أن نحلل و ولماذا نريد القيام بتحليل هذه البيانات .. أي ما هو الغرض أو الهدف من إجراء التحليل تم ننتقل لتحديد نوع تحليل البيانات المراد استخدامه , ثم التفكير في الطريقة المراد قياس بها أي تغيير نحدثه على البيانات .

2– جمع البيانات :

بعد خطوة جمع متطلبات البيانات , نحصل على فكرة واضحة عن المتغيرات المراد قياسها وماذا يجب أن تكون النتائج التي توصلنا إليها.

لننتقل إلى خطوة الثانية وهي جمع البيانات بناء ً على المتطلبات.

وبمجرد جمع البيانات ، يجب معالجة البيانات التي تم جمعها أو تنظيمها للتحليل. أثناء جمع البيانات من مصادر مختلفة , يجب أن نحتفظ بسجل مع تاريخ جمع ومصدر للبيانات.

3– تنظيف البيانات :

مهما كانت البيانات التي تم جمعها قد لا تكون مفيدة أو غير واضحة أو غير ذات صلة بهدف التحليل الخاص بالبحث , وبالتالي يجب تنظيفها.

فقد تحتوي البيانات التي يتم جمعها على سجلات مكررة أو مسافات بيضاء أو أخطاء أو حتى قيم تحتاج إلى توضيح أو جمع ...إلخ وعليه يجب تنظيف البيانات لتكون خالية من الأخطاء.

يجب أن يتم تنفيذ هذه المرحلة قبل التحليل لأنه بناء ً على تنظيف البيانات ، سيكون ناتج التحليل أقرب إلى النتيجة المتوقعة.

4- تحليل البيانات :

بمجرد جمع البيانات وتنظيفها ومعالجتها ، تصبح جاهزة للتحليل.

أثناء معالجة البيانات , قد نجد أن لدينا المعلومات الدقيقة التي نحتاجها , أو قد نحتاج إلى جمع المزيد من البيانات. خلال هذه المرحلة ، يمكننا استخدام أدوات وبرامج تحليل البيانات التي ستساعدك على فهم الاستنتاجات وتفسيرها واستنتاجها بناءً على المتطلبات.

5- تفسير البيانات :

بعد تحليل بياناتك ، ننتقل إلى خطوة تفسير النتائج , حيث يمكننا اختيار طريقة لتمثيل النتائج و إيصال المخرجات إلى الجهات المعنية وذلك إما باستخدام نموذج الكلمات (السرد) أو نموذج جدول البيانات أو حتى المخططات , ثم يتم استخدم نتائج عملية تحليل البيانات لتحديد أفضل مسار للعمل.

6- عرض مرئي للمعلومات :

الخطوة الأخيرة من خطوات عملية تحليل البيانات

وهي إظهار البيانات بيانيًا بحيث يسهل على الدماغ البشري فهمها ومعالجتها.

غالبًا ما يستخدم تصور البيانات لاكتشاف الحقائق والاتجاهات غير المعروفة , وذلك من خلال مراقبة العلاقات ومقارنة مجموعات البيانات , يمكننا العثور على طريقة لاكتشاف معلومات مفيدة.

تقنية التنقيب في المعطيات – Data Mining

تعريف:

شهدت تقنية التنقيب في المعطيات خطوات سريعة على مدى العقدين الماضيين , وخاصة من وجهة نظر مجتمع علوم الكمبيوتر.

وذلك نتيجة لطوفان البيانات الناتج بشكل مباشر للتقدم في التكنولوجيا والحوسبة

بينما تم دراسة تحليل البيانات على نطاق واسع في المجال التقليدي للاحتمالات والإحصائيات.

حيث يعد استخراج البيانات مصطلحًا صاغهُ مجتمع موجه نحو علوم الكمبيوتر , حيث أن لعلماء الكمبيوتر قضايا مثل قابلية التوسع والاستخدام والتنفيذ الحسابي , إلى أن ظهر مصطلح علم البيانات الذي كان كسمة مشتركة مع علم الكمبيوتر متخصصًا بذاته بدراسة مجال البيانات واستخلاص المعرفة منها , ليكون التنقيب في المعطيات تحت مظلته كتقانة من تقنياته ضمن مجاله الواسع .

التنقيب في المعطيات هو دراسة جمع وتنظيف ومعالجة وتحليل واكتساب فائدة و رؤى من البيانات.

في العصر الحديث , تولد جميع المنظومات الآلية تقريبًا شكلًا من أشكال البيانات أيضًا لأغراض التشخيص أو التحليل , وقد أدى ذلك إلى طوفان في البيانات ، والذي تم الوصول إلى ترتيب البيتابايت أو إكسابايتس في وحدات حجم البيانات.

الأنواع البيانات التي يمكن التنقيب عليها:

كتقنية عامة ، يمكن تطبيق التنقيب البيانات على أي نوع من البيانات طالما أن

البيانات ذات مغزى للتطبيق المستهدف , أبسط أشكال البيانات للتنقيب هي بيانات قاعدة البيانات – بيانات مستودع البيانات – وبيانات المعاملات – البيانات المرتبة

الرسم البياني أو البيانات المتصلة بالشبكة – وبيانات الوسائل المتعددة **www**

مثال:

1 – World Wide Web :

عدد المستندات الموجودة على الويب المفهرس وموجود الآن في الطلب تقدر بالمليارات هذا في قياسنا التقريبي لما يحتويه الويب المرئي متناسين ما خفي في **deep web**

ليتم استثمار التنقي عن المعطيات طريق تحديد الارتباطات بين مواضيع مختلفة على الويب , أو من ناحية أخرى يمكن لسجلات وصول المستخدم أن يتم استخراجه لتحديد

أنماط الوصول المتكررة أو أنماط غير عادية من المحتمل سلوك غير مبرر.

2- التفاعلات المالية:

معظم المعاملات الشائعة في الحياة اليومية , مثل استخدام بطاقة الصراف الآلي (ATM) أو بطاقة الائتمان , تنشئ بيانات آلية يمكن تطبيق خوارزميات التنقيب في المعطيات للحصول على العديد من الأفكار المفيدة مثل النشاطات الغير مرغوبة كالأحتيال.

3- فاعلات المستخدم :

تنشئ العديد من أشكال تفاعلات المستخدم كميات كبيرة من البيانات كاستخدام الهاتف الذي بدوره ينشئ ما يسمى بسجل الاتصالات .. ليتم استثمار هكذا نوع من البيانات في لاتخاذ قرارات بشأن التسعير و سعة الشبكة أو حتى استهداف العملاء

عملية التنقيب في البيانات \ The Data Mining Process

كما تم الشرح من قبل عن عملية تحليل البيانات سيتم إستئناف حديثنا في توضيح ماهية عملية التنقيب في المعطيات , فهي كمثيلتها في تحليل البيانات تحتوي على العديد من الخطوات مثل تنظيف البيانات واستخراج المعرفة والتصميم الخوارزمي

1- جمع البيانات \ Data collection

يتطلب جمع البيانات استخدام أجهزة متخصصة مثل , شبكة أجهزة الاستشعار أو عن طريق العمل اليدوي كأستطلاعات المستخدمين أو عن طريق أدوات برمجية للنقل بعد عملية جمع البيانات إلى تخزينها في ملفات قاعدة البيانات أو حتى مستودع لمعالجة البيانات.

2- استخراج الميزات و تنظيف البيانات \ Feature extraction and data cleaning

عندما قام الباحث بعملية جمع البيانات فغالب الأمر أن لا يتم جمعها على الشكل المناسب للمعالجة , حيث من الممكن أن تتواجد مجموعة كبيرة من البيانات المجمعة والمخزنة بشكل تعسفي , وعليه في هذه الخطوة المهمة الأساسية للباحث هو تحويل هذه البيانات إلى الشكل و التنسيق المناسبين للمعالجة , وبعد مرحلة التنظيف يتم العودة مرة لتخزين البيانات في قاعدة البيانات.

3- اختيار البيانات \ Data selection

حيث يتم استرداد البيانات ذات الصلة بمهمة التحليل من قاعدة البيانات

4- تحويل البيانات \ Data transformation

حيث يتم تحويل البيانات ودمجها في أشكال مناسب للتنقيب .

5- استخراج البيانات \ Data mining

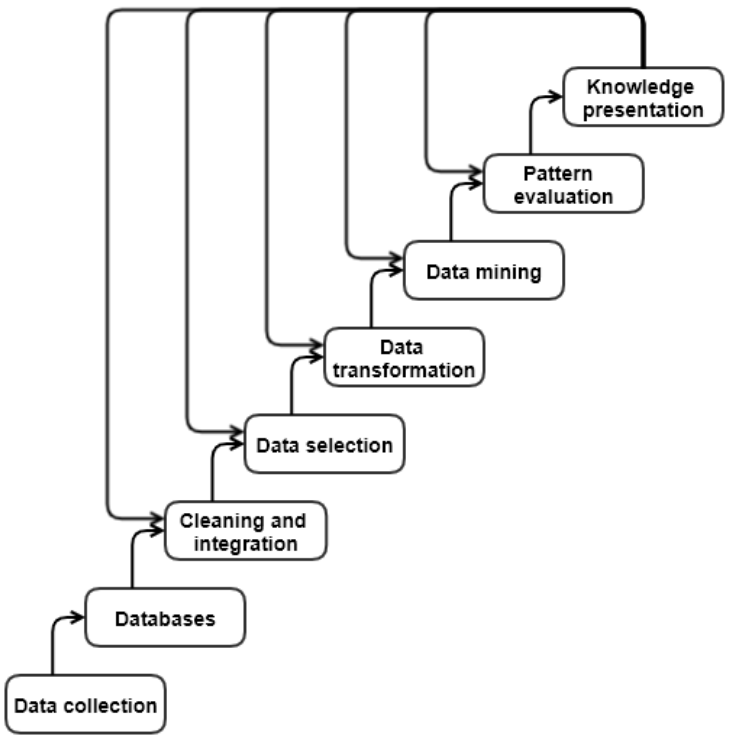
عملية أساسية حيث يتم تطبيق أساليب ذكية لاستخراج أنماط البيانات

6- تقييم الأنماط \ Pattern evaluation

وذلك لتحديد الأنماط المثيرة للاهتمام في البيانات التي تمثل المعرفة أو قيم مضافة إلى المنظومة بناء

7- عرض المعرفة \ Knowledge presentation

حيث تقنيات التصوير وتمثيل المعرفة تستخدم لتقديم المعرفة الملغومة للمستخدمين



The data processing pipeline

أنواع أنماط التنقيب :

Data characterization
Data discrimination
Associations Analysis
Classification for Predictive Analysis
Cluster Analysis
Outlier Analysis

Data characterization :

إن عملية توصيف البيانات تتمثل بتلخيص للميزات العامة للكائنات في فئة مستهدفة , وينتج ما يسمى قواعد الخصائص الخاصة بهذا الكائن , أي كما هو حال المريض الذي يذهب بانتظام إلى العيادة الطبية بشكل شهري فيأتي توصيف البيانات من خلال التنقيب في خصائص هذا الشخص وأمثاله حول كم ينفقون شهرياً .. ما هي وامفاتهم ..وما هو الوضع المالي لهم

Data discrimination :

يُنتج تمييز البيانات ما يسمى القواعد التمييزية وهو في الأساس مقارنة بين السمات العامة للأشياء بين فئتين, يشار إليهما بالفئة المستهدفة والفئة المتناقضة.

أي كما هو الحال عند مقارنة الخصائص العامة للمرضى الذين زاروا عيادة الطبيب أكثر من 10 زيارات في العام الماضي مع أولئك الذين يقل حسابهم في الزيارات عن 2.

و كملاحظة فإن التقنيات المستخدمة لتمييز البيانات تشبه إلى حد كبير الأساليب المستخدمة لتوصيف البيانات باستثناء أن نتائج تمييز البيانات تشمل مقاييس مقارنة.

: Associations Analysis

هو اكتشاف ما يُعرف باسم قواعد الارتباط.

حيث يقوم بدراسة تواتر العناصر التي تحدث معًا في قواعد بيانات المعاملات ، وعلى أساس عتبة تسمى الدعم (support) ، يحدد مجموعات العناصر المتكررة.

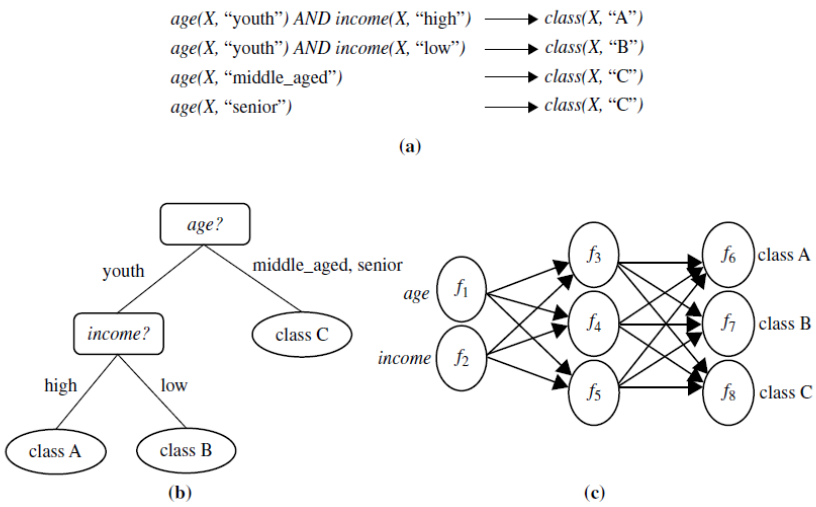
عتبة الثقة (Confidence) وهي الاحتمال الشرطي من ظهور عنصر في معاملة عندما يظهر عنصر آخر ، يتم استخدامها لتحديد قواعد الاقتران.

: Classification

التصنيف: تحليل التصنيف هو تنظيم البيانات في فئات معينة.

يُعرف التصنيف أيضًا باسم التصنيف الخاضع للإشراف ، ويستخدم خوارزميات تصنيفية معينة لترتيب الكائنات في مجموعة البيانات.

تستخدم خوارزمية التصنيف في مجموعة التدريب وتبني نموذجًا , يستخدم النموذج لتصنيف الكائنات الجديدة.



يمكن تمثيل نموذج التصنيف بأشكال مختلفة:

(a) IF-THEN rules	(b) a decision tree	(c) neural network

: Cluster Analysis

العنقدة: على غرار التصنيف ، العنقدة هي عملية تنظيم البيانات في فئات.

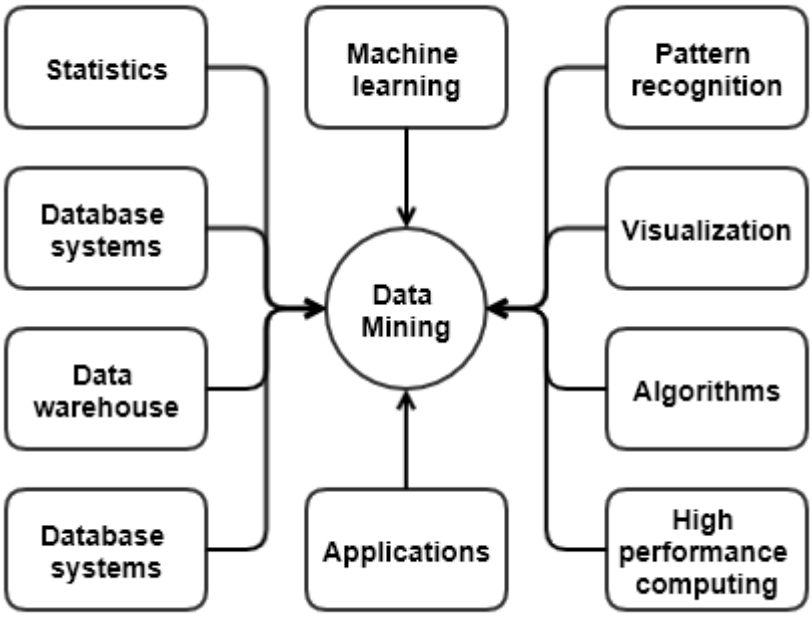
ومع ذلك ، على عكس التصنيف ، في تصنيفات المجموعات ، تكون تسميات الصف غير معروفة ويعود الأمر إلى خوارزمية التجميع لاكتشاف فئات مقبولة.

يسمى التجميع أيضًا التصنيف غير الخاضع للإشراف , لأن التصنيف لا تمليه تسميات فئة معينة. هناك العديد من مقاربات التجميع تعتمد جميعها على مبدأ زيادة التشابه بين الكائنات في نفس الفئة (التشابه داخل الفئة) وتقليل التشابه بين كائنات الفئات المختلفة (التشابه بين الفئات).

: Outlier Analysis

إن القيم المتطرفة هي عناصر بيانات لا يمكن تجميعها في فئة أو مجموعة معينة, وغالبًا ما يكون من المهم جدًا تحديدها, حيث أن القيم المتطرفة يمكن اعتبارها ضوضاء ويجب التخلص منها في بعض التطبيقات ، إلا أنها يمكن أن تكشف عن معرفة مهمة في مجالات أخرى ، وبالتالي يمكن أن تكون مهمة للغاية ويكون تحليلها ذا قيمة

التقنيات المستخدمة في التنقيب عن المعطيات:



مقدمة:

كما ذكرنا في الفصول الماضية حيث أن البيانات هي النفط الجديد في عصرنا الحالي.

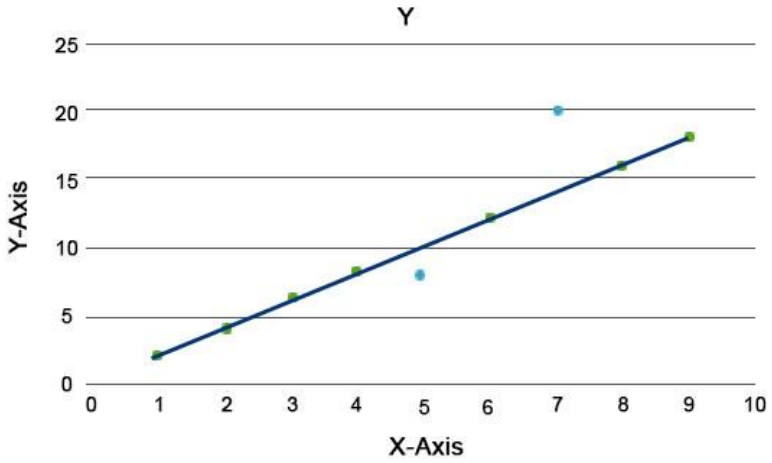
يجدر الذكر إلى أنه تختلف نتيجة مشروع علم البيانات اختلافًا كبيرًا مع نوع البيانات المتاحة , نظرًا لوجود العديد من أنواع التحليلات المختلفة المتاحة , يصبح من الضروري فهم بعض تقنيات خطوط الأساس التي يجب اختيارها.

حيث أن الهدف الأساسي لتقنيات علوم البيانات ليس فقط البحث عن المعلومات ذات الصلة ولكن أيضًا اكتشاف الروابط الضعيفة التي تميل إلى جعل النموذج يعمل بشكل سيئ.

أنواع تقنيات التنقيب في المعطيات :

- **UnSupervised Learning**

:Anomaly Detection



النقاط الزرقاء هي النقاط الشاذة في النموذج مبتعدة عن الإتجاه المستقيم الأساس

إن أبسط نهج لتحديد النقاط المخالفة (أحداث غير متوقعة / إنحرافات) في مجموعات البيانات هو وضع علامة على نقاط البيانات التي تنحرف عن الخصائص الإحصائية الشائعة للتوزيع , بما في ذلك المتوسط والوسيط والنمط والكميات.

حيث نفترض أن تعريف نقطة البيانات الشاذة هو انحراف عن طريق انحراف معياري معين عن المتوسط.

سنحتاج إلى متغير متحرك لحساب المتوسط عبر نقاط البيانات. من الناحية الفنية ، يُسمى هذا المتوسط المتداول أو المتوسط المتحرك ، ويقصد به تخفيف التقلبات قصيرة المدى وإبراز التقلبات طويلة المدى.

حيث أن للكشف عن الشذوذ افتراضان أساسيان:

نادرًا ما تحدث الشذوذات في البيانات.

تختلف مميزاتها عن الحالات العادية بشكل كبير.

وإن أهم تقنيات الكشف عن الشذوذ هي الأساليب الإحصائية البسيطة

Association Analysis :

يساعد هذا التحليل في بناء علاقات مثيرة للاهتمام بين العناصر في مجموعة بيانات.

حيث يكشف عن العلاقات الخفية ويساعد في تمثيل عناصر مجموعة البيانات في شكل قواعد الارتباط أو مجموعات العناصر المتكررة. يتم تقسيم قاعدة الارتباط إلى خطوتين:

1 – إنشاء مجموعة عناصر متكررة: أي يتم إنشاء مجموعة حيث يتم إعداد العناصر المتكررة معًا.

2– إنشاء القاعدة: يتم تمرير المجموعة التي تم بناؤها في الخطوة (1) من خلال طبقات مختلفة من تكوين القواعد لبناء علاقة مخفية فيما بينها.

Clustering Analysis :

سيشرح بشكل معمق أدناه في قسم التنقيب في المعطيات

- **Supervised Learning**

Regression Analysis :

تحليل الانحدار هي تقنية تساعد في فهم العلاقات بين المتغير المستقل الذي يساعد في تحديد علاقته بالمتغير التابع الذي يؤثر على متغير النتيجة (المتغير الهدف أو المتغير التابع) , وتحليل النتيجة بتغير المتغير المستقل باستخدام تقنيات الإنحدار المختلفة مثل Linear Regression (يستخدم مبدأ الطريقة المربعات الصغرى) ، Logistic Regression و Lasso/Ridge Regression

Modeling Logistic Growth :

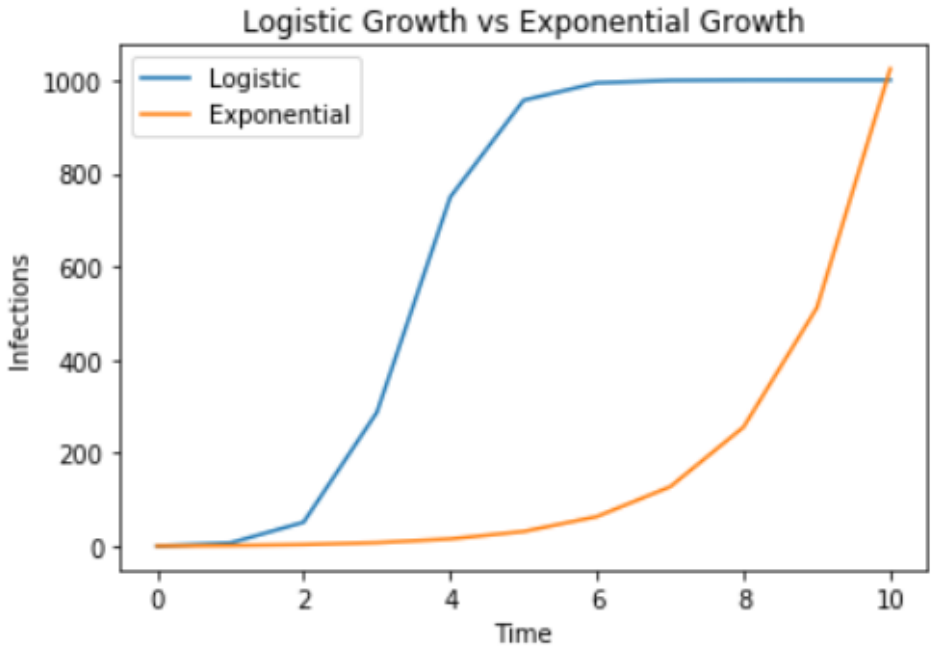
النمو اللوجستي هو تقنية رياضية يمكن استخدامها في العديد من الحالات , حيث يتميز النمو اللوجستي بزيادة النمو في فترة البداية ، ولكن النمو يتناقص في مرحلة لاحقة وذلك كلما تم الإقتراب من الحد الأقصى.

(سيعطي مخططه المرئي تصورا عن وصول متغير ما مثلا إلى حده (حد الذروة))

في حالات الاستخدام الأخرى للنمو اللوجستي ، يمكن أن يكون هذا الرقم هو حجم تعداد الإصابات الذي ينمو بشكل كبير حتى اللحظة التي لا يمكن فيها إنتشار العدوى كحالها في السابق , وبالتالي يصبح النمو أبطأ حتى يتم الوصول إلى الحد الأقصى لسعة الإصابات .

ملاحظة:

سبب استخدام النمو اللوجستي لنمذجة تفشي الفيروس التاجي (COVID_19) هو أن علماء الأوبئة درسوا هذه الأنواع من الفاشيات ومن المعروف جيدًا أن الفترة الأولى للوباء تتبع النمو الأسي وأن الفترة الإجمالية اللاحقة يمكن نمذجتها بالنمو اللوجستي (قمنا بتطبيق ذلك في القسم العملي للبحث)



يتميز النمو اللوجستي بالصيغة الرياضية التالية:

$$y(t) = \frac{c}{1 + a * e^{-bt}}$$

بحيث أن Y (t) هو عدد الحالات في أي وقت , وأن C هي القيمة الحدية ، السعة القصوى لـ y

وأيضاً

عدد الحالات في البداية ، وتسمى أيضاً القيمة الأولية: c / (1 + a)

الحد الأقصى لمعدل النمو هو t = ln (a) / b و y (t) = c / 2

لتوضيح ما ذكر بطريقة رياضية أكثر تفصيلاً

يتم اشتقاق الوظيفة أعلاه في الواقع من الصيغة التفاضلية أدناه التي اكتشفها بيير فرانسوا فيرهولست.

$$\frac{dy}{dt} = ry * (1 - \frac{y}{c})$$

حيث يشير الجزء الأخضر dy / dt إلى أن هذه الصيغة ليست لحجم السكان ، ولكن لنمو السكان.

كما يمكننا أن نرى أن y و c في الصيغة , لذلك نفهم أن نمو السكان يعتمد على قيمة y (حجم السكان) وقيمة c (السعة القصوى)

عندما تكون y مساوية لـ c (أي أن الحجم الأقصى للسكان) ، فإن y / c ستكون (1) وبالتالي ، سيكون الجزء الأزرق ٠ وبالتالي يكون النمو ٠.

عندما تكون y أصغر بكثير من c (يكون السكان بعيداً عن الحد) سيكون الجزء الأزرق تقريباً (1). لذلك ، يتم تعريف النمو بواسطة الجزء البرتقالي الذي هو في الواقع صيغة للنمو الأسي.

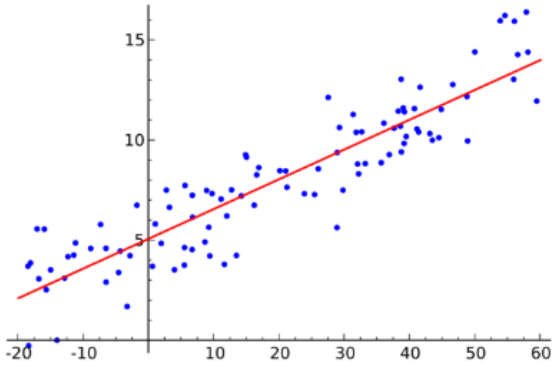
Classification Analysis :

سيشرح بشكل معمق أدناه في قسم التنقيب في المعطيات

Data Science Algorithms

Type 1: Supervised Algorithms.

Type 2: Unsupervised Algorithms.



علم البيانات هو مجال للدراسة حيث يتم اتخاذ القرارات بناء على الرؤى التي نحصل عليها من البيانات بدلا من الأساليب الحتمية الكلاسيكية القائمة على القواعد.

في دورة الحياة الخاصة بعلم البيانات, نستخدم خوارزميات علم البيانات المختلفة لحل المهام قيد التنفيذ

أنواع خوارزميات علوم البيانات :

Supervised Algorithms

Unsupervised Algorithms

Supervised Algorithms VS Unsupervised Algorithms

Supervised Algorithms: عندما يكون لدينا متغيرات الإدخال ومتغير الإخراج وتستخدم خوارزمية لتعلم وظيفة رسم الخرائط من الإدخال إلى الإخراج. والهدف من ذلك هو تقريب وظيفة التعيين بحيث عندما يكون لدينا بيانات إدخال جديدة يمكننا التنبؤ بمتغيرات الإخراج لتلك البيانات.

Unsupervised Algorithms: هو نمذجة البنية الأساسية أو المخفية أو التوزيع في البيانات من أجل معرفة المزيد عن البيانات. التعلم غير الخاضع للرقابة هو المكان الذي لديك فيه بيانات الإدخال فقط وليس لديك متغيرات مخرجات مقابلة.

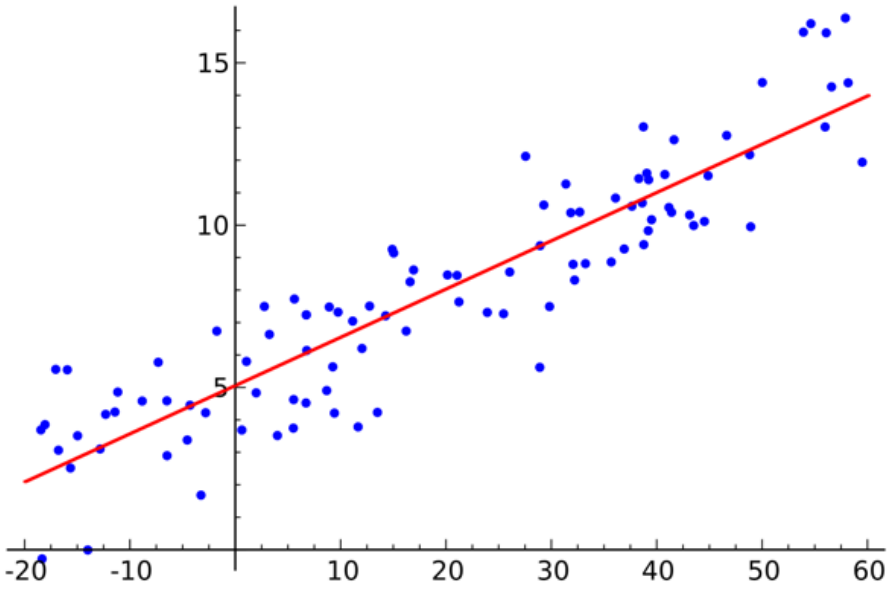
- **Supervised Algorithms** :

1- K Nearest Neighbors :

(KNN) هي واحدة من أبسط خوارزميات التعلم الآلي المستخدمة في علم البيانات ذو قدرات عمل عالية , وهي خوارزمية خاضعة للإشراف حيث يتم التصنيف بناءً على أقرب نقاط بيانات.

تتمثل الفكرة وراء KNN في تجميع نقاط متشابهة معًا , من خلال قياس خصائص أقرب نقاط بيانات يمكننا تصنيف نقطة بيانات اختبار.

2- Linear Regression :



هي نوع من خوارزمية تعلُم الآلة تُستخدم لنمذجة العلاقة بين التبعية العددية ومتغير مستقل واحد أو أكثر هو إحدى طرق إجراء التحليل التنبؤي.

للتنبؤ بالنتيجة من مجموعة متغيرات التنبؤ , ما المتغيرات التنبؤية التي لها أقصى تأثير على متغير النتيجة؟

تُعرف حالة وجود متغير مستقل واحد باسم الانحدار الخطي البسيط بينما تُعرف حالة وجود أكثر من متغير مستقل الانحدار الخطي المتعدد. في كل من هذه الانحدارات الخطية

صيغة الانحدار الخطي: $y = mx + c$

حيث ان

x: نقاط المتغير المستقل

M: معامل الانحدار

C: ثابت

استخدامات الانحدار الخطي :

يستخدم الانحدار الخطي لإجراء تحليل الانحدار. فيما يلي استخدامات تحليل الانحدار.

١- يساعد في تحديد قوة المتنبئين: يساعد التحليل التنبؤي في فهم العلاقة بين المتنبئ ومتغير النتيجة (أي الجرعة والتأثير).

٢- توقع التأثير من خلال التنبؤ: سيؤدي التغيير في المتغير التابع إلى اختلاف في متغير مستقل.

٣- تحليل / التنبؤ بالاتجاه: يستخدم تحليل الانحدار للتنبؤ بالاتجاهات المستقبلية خاصة في سوق الأسهم حيث توجد تقلبات وتضخم في الأسعار.

أنواع الانحدار الخطي :

فيما يلي أنواع الانحدار الخطي:

Simple Linear Regression :

يحتوي الانحدار البسيط على متغير تابع واحد (فاصل زمني أو نسبة) ومتغير واحد مستقل (فاصل زمني أو نسبة أو ثنائي التفرع).

Multiple Linear Regression

يستخدم الانحدار المتعدد عندما يكون لدينا متغيرين مستقلين ومتغير تابع واحد. يمكننا تحديد تأثير المتغيرات المستقلة على المتغير التابع.

ومنه يمكننا أن نفترض أن **x** تكون سلسلة من المتغيرات المستقلة (x1، x2...) و Y لتكون متغيراً تابعاً. لدينا أيضاً **b** كمنحدر لمتغير الانحدار.

ومنه تكون المعادلة التي تمثل العلاقة بين x وy.

$$Y = a + b1x1 + b2x2 + \dots$$

Ordinal Regression :

يتم إجراء الانحدار العادي على متغير ثنائي التفرع تابع ومتغير مستقل واحد يمكن أن يكون ترتيبيًا أو اسميًا. يمكن إجراء الانحدار العادي باستخدام النموذج الخطي المعمم (GLM)

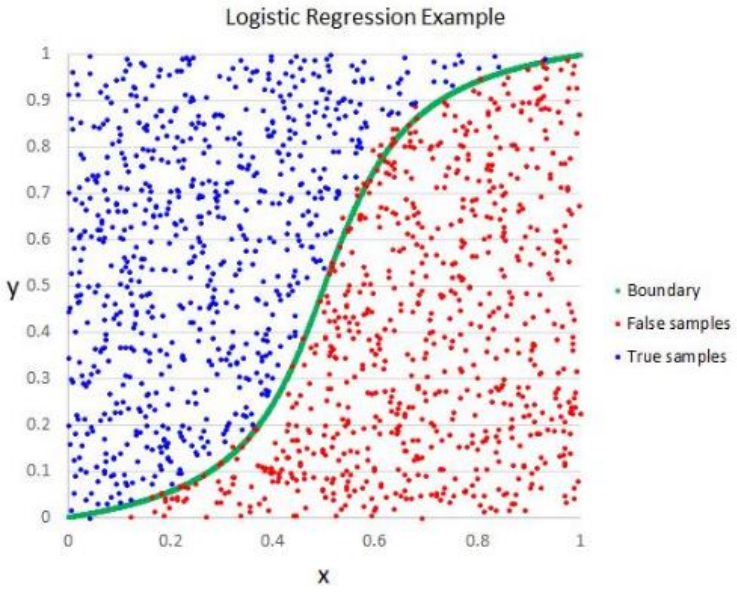
كمثال : التنبؤ بما إذا كان شراء المنتج يمكن أن يؤدي إلى قيام المستهلك بشراء منتج ذي صلة

Multinonial Regression :

يتم إجراء الانحدار متعدد الحدود على متغير تابع اسمي ومتغير مستقل واحد وهو النسبة أو الفاصل الزمني أو ثنائي التقسيم.

أي يمكن أن يكون كالتفضيلات المهنية بين الطلاب الذين يعتمدون على مهنة الوالدين وتعليمهم.

Logistic Regression :



على الرغم من أن الاسم يشير إلى الانحدار , فإن الانحدار اللوجستي هو خوارزمية تصنيف خاضعة للإشراف.

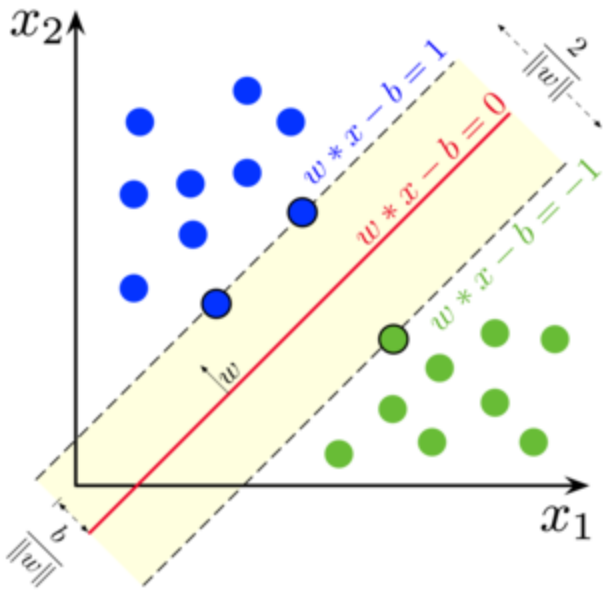
يتم الانحدار اللوجستي عندما يكون هناك متغير تابع واحد ومتغيرين مستقلين, الفرق بين الانحدار المتعدد واللوجستي هو أن المتغير الهدف منفصل (قيمة ثنائية أو قيمة ترتيبية).

إن مشكلة الانحدار الخطي هي أن القيمة المتغيرة ثابتة فقط على نتيجتين محتملتين.

من ناحية أخرى ، يمكن أن يرجع الانحدار اللوجستي درجة الاحتمالية التي تنعكس على حدوث حدث معين.

يمكن أن يستخدم في الكشف عن رسائل البريد الإلكتروني العشوائية , والتنبؤ بمبلغ قرض العميل.

Support Vector Machine -3 :



يمكن استخدام هذا النوع من الإنحدار , المختصر باسم SVM في كل من مهام الانحدار والتصنيف. ولكن , يتم استخدامه على نطاق واسع في أهداف التصنيف.

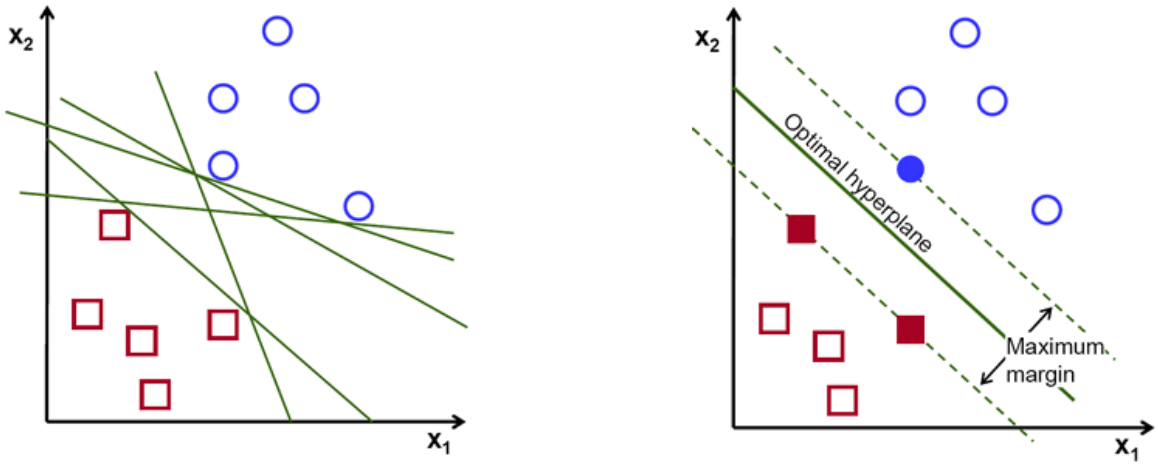
حيث يقوم هذا الإنحدار ببناء نموذج يعين أمثلة جديدة لفئة واحدة أو أخرى ، مما يجعلها مصنفا خطي ثنائي غير احتمالي , حيث يعمل على تمثيل الأمثلة كنقاط في الفضاء ، بحيث يتم تقسيم أمثلة الفئات المنفصلة على فجوة واضحة واسعة قدر الإمكان. بعد ذلك يتم تعيين أمثلة جديدة في نفس المساحة ويتوقع أن تنتمي إلى فئة تستند إلى جانب الفجوة التي تقع عليها.

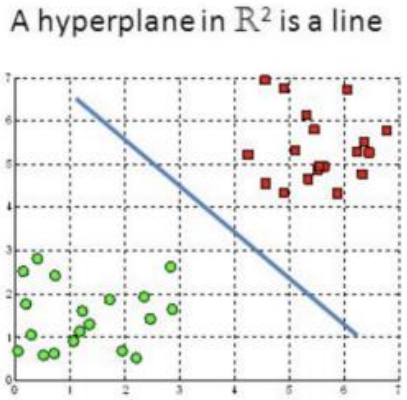
بالإضافة إلى إجراء التصنيف الخطي ، يمكن لـ SVM إجراء تصنيف غير خطي بكفاءة باستخدام ما يسمى kernel trick ، وتعيين مدخلاتهم بشكل ضمني في مساحات الميزة عالية الأبعاد.

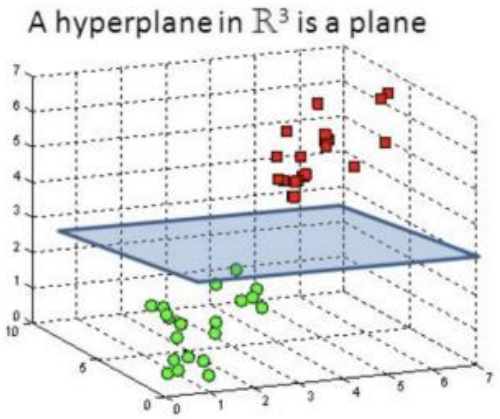
ملاحظة: حول hyperplane

في الهندسة , فإن hyperplane هي مساحة فرعية يكون أبعادها أقل من أبعاد المساحة المحيطة بها.

إذا كانت المساحة ثلاثية الأبعاد ، فإن hyperplane هي ثنائية الأبعاد ، بينما إذا كانت المساحة ثنائية الأبعاد ، فإن هي خطوط أحادية البعد. يمكن استخدام هذه الفكرة في أي مساحة عامة يتم فيها تحديد مفهوم البعد من الفضاء الجزئي.







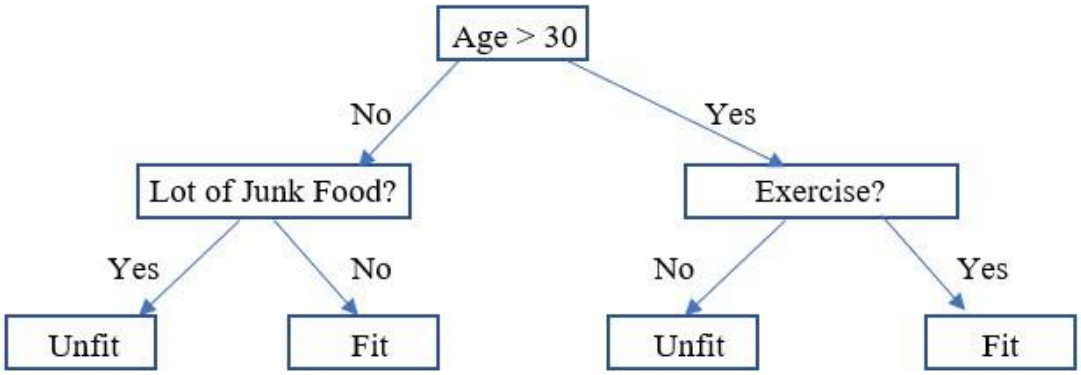
4- Decision Tree :

شجرة القرار هي الخوارزمية الهيكلية تستخدم للحصول على مخرجات ذات مغزى من مجموعة متنوعة من المدخلات. يعتبر الناتج الذي يتم الحصول عليه من هذا النوع من الترتيب الهرمي مساهمة قيمة لإنتاج نتائج تحليلية لاتخاذ القرارات التجارية الأساسية. ومن هنا جاء اسم شجرة القرار لهذه الخوارزمية.

حيث أن شجرة القرار هي خوارزمية تصنع بنية تشبه الشجرة أو هيكل يشبه المخطط الانسيابي حيث يكون كل مستوى أو ما نسميه العقدة بمثابة اختبار يعمل على ميزة.

وإن في عقدة البداية , تقع مجموعة البيانات بأكملها في فئة واحدة , وبعد ذلك في كل طبقة ، يتم تقسيم البيانات.

Age	A lot of Junk food	Exercise	Fit/Unfit
45	N	N	Unfit
40	N	Y	Fit
25	Y	N	Unfit
29	N	Y	Fit
23	Y	Y	Unfit



مميزات شجرة القرارات انه يتم تطبيقها على التحليلات التنبؤية لحل المشاكل وتستخدم في الأنشطة اليومية لاختيار الهدف على أساس تحليل القرار. تقوم تلقائيًا ببناء نموذج بناءً على بيانات المصدر. الأفضل في التعامل مع القيم المفقودة.

أما من عيوبها فإن حجم الشجرة لا يمكن السيطرة عليه حتى لديها بعض معايير التوقف. بسبب شجرة هيكلها الهرمي غير مستقر.

- **Unsupervised Algorithms**

1- K Means++

هي خوارزمية لاختيار القيم الأولية (أو "الأصول \ البذور") لخوارزمية التجميع

تم اقتراحها في عام 2007 من قبل **David Arthur** وسيرجي فاسيلفيتسكي ، كخوارزمية تقريب لمشكلة الوسائل NP-hard – وهي طريقة لتجنب التجمعات الضعيفة في بعض الأحيان التي تم العثور عليها بواسطة خوارزمية k means.

حيث تعتمد خوارزمية k means بشكل كبير على تهيئة النقط الوسطى.

والذي بدوره سينتج عن التهيئة مجموعات ضعيفة.

وعليه تم تصميم K-mean++ أتت لتحسين تهيئة النقطه الوسطى للوسائل , حيث أن الفكرة الأساسية هي أن النقطه الوسطى الأولية يجب أن تكون بعيدة عن بعضها البعض.

تبدأ الخوارزمية باختيار النقطة الوسطى بشكل عشوائي من جميع نقاط البيانات.
بالنسبة إلى centroid ci ، بهذه الطريقة تحاول k-mean ++ دائمًا اختيار النقط الوسطى البعيدة عن النقط الوسطى الحالية ، مما يؤدي إلى تحسن كبير على حساب التضحية في وقت التشغيل.

K Medoids -2

هي خوارزمية تجميع مرتبطة بخوارزمية k-mean وخوارزمية medoidshift.

تحاول K-mean تقليل إجمالي الخطأ التربيعي , بينما تقلل k-medoids مجموع الاختلافات بين النقاط المصنفة في مجموعة ونقطة محددة كمركز لتلك المجموعة.

K Means -3

سيشرح بشكل معمق أدناه في قسم التنقيب في المعطيات

1- Descriptive Analysis - التحليل الوصفي

يأخذ التحليل الوصفي في الاعتبار البيانات التاريخية ، ومؤشرات الأداء الرئيسية ، ويصف الأداء بناء ً على المعيار المختار. يأخذ في الاعتبار الاتجاهات السابقة وكيف يمكن أن تؤثر على الأداء في المستقبل

2- Dispersion Analysis - تحليل التشتت

التشتت في المنطقة التي تنتشر فيها مجموعة البيانات. تسمح هذه التقنية لمحللي البيانات بتحديد تنوع العوامل قيد الدراسة.

3– Factor Analysis - تحليل العوامل

يساعد هذا الأسلوب على تحديد ما إذا كانت هناك أي علاقة بين مجموعة من المتغيرات.

في هذه العملية ، يكشف عن عوامل أو متغيرات أخرى تصف أنماط العلاقة بين المتغيرات الأصلية. يقفز تحليل العوامل إلى الأمام في إجراءات التجميع والتصنيف المفيدة.

4– Discriminant Analysis - التحليل التمييزي

هي تقنية تصنيف في استخراج البيانات. تحدد النقاط المختلفة في المجموعات المختلفة بناء ً على القياسات المتغيرة. بعبارات بسيطة ، تحدد ما يجعل مجموعتين مختلفتين عن بعضهما البعض ؛ هذا يساعد على تحديد عناصر جديدة.

5– time-series - تحليل السلاسل الزمنية

تسمى سلسلة نقاط البيانات المسجلة خلال فترة زمنية محددة ببيانات سلسلة زمنية.

حيث أن تحليل السلاسل الزمنية هو تقنية لتحليل بيانات الزمنية واستخراج معلومات إحصائية وخصائص ذات مغزى للبيانات , أحد الأهداف الرئيسية للتحليل هو التنبؤ بالقيمة المستقبلية , حيث يتم إجراء الاستقراء عند التنبؤ بتحليل السلاسل الزمنية وهو أمر معقد للغاية.

ولكن ، القيمة المتوقعة مع تقدير عدم اليقين المرتبط بذلك يمكن أن تجعل النتيجة ذات قيمة للغاية.

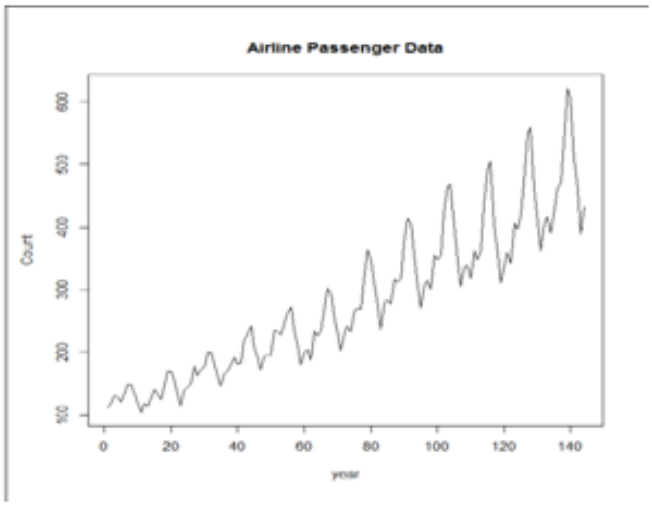
ما هو تحليل السلاسل الزمنية؟

ومنه نستنتج بأنها طريقة إحصائية لتحليل البيانات السابقة في غضون فترة زمنية معينة للتنبؤ بالمستقبل.

تتألف من تسلسل مرتب للبيانات على مسافات متساوية ، لفهم بيانات السلاسل

يبدأ تحليل السلسلة الزمنية بالتحليل الاستكشافي Exploratory Analysis

الذي يتم عن طريق رسم مخطط خطي لعدد النقاط مع مرور الوقت.



الشكل(١)

حيث يوضح الشكل(ا) أعلاه عدد متغير ما على المحور Y والوقت على المحور X

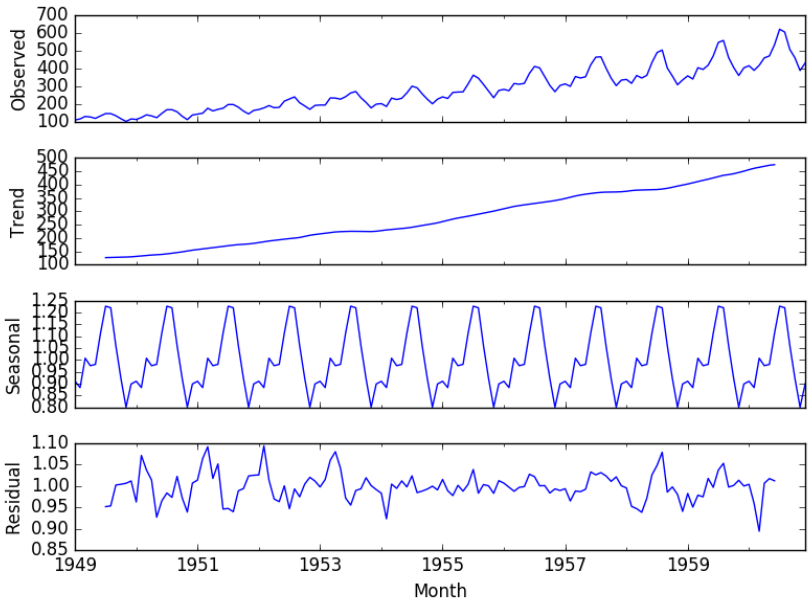
حيث يمكن أن يشتق من المخطط أعلاه الملاحظات التالية:

الاتجاه: TREND وهو نمط زيادة أو تناقص على مدى فترة من الزمن. كما يبين لنا في المثال , لوحظ الاتجاه الأساسي المتزايد تدريجيًا. أي ازداد عدد المتغير على مدى فترة من الزمن.

الموسمية: SEASONALITY وهو النمط الدوري للبيانات.

نمط مماثل يتكرر بعد فترة زمنية معينة. كما هو ملاحظ في المتغير أعلاه نمط دوري يحتوي على ارتفاع معين ونقطة منخفضة يمكن رؤيته في جميع الفترات.

فرط التباين: HETROSCEDASTICITY وهو التباين الغير ثابت أو انحراف متغير عن المتوسط على مدى فترة من الزمن. في الرسم البياني أعلاه ازداد التباين باستمرار على مدار فترة من الزمن.



في السلسلة الزمنية ، الوقت هو المتغير المستقل والهدف هو التنبؤ.

أسباب الرئيسية لإجراء تحليل السلاسل الزمنية:

لتفسير الميزات: ميزات مثل الاتجاه (زيادة أو نقصان) ، الموسمية (النمط الدوري) والتنوع (المرونة غير المتجانسة)

التنبؤ: يعد هذا أحد أهم أسباب إجراء تحليل السلاسل الزمنية. على سبيل المثال ، يمكن أن تساعدنا السلاسل الزمنية في التنبؤ بسعر السهم. لأننا لن نهتم فقط بمعرفة ما إذا كان السعر سيرتفع أو ينخفض ولكن أيضًا المبلغ الذي سيرتفع به أو ينخفض.

الاستدلالات: كما نعلم جميعًا ، من الصعب التنبؤ بالمستقبل الدقيق. لذا يمكن التعامل مع القيمة المتوقعة جنبًا إلى جنب مع هامش الخطأ (عدم اليقين) كنتيجة أفضل. وبمساعدة تحليل السلاسل الزمنية ، يمكننا استخلاص الاستنتاج مثل فاصل الثقة واختبار الفرضية على معلمات النموذج

السلاسل الزمنية للنمذجة :

هناك طرق عديدة لصياغة سلسلة زمنية من أجل عمل تنبؤات:

المتوسط المتحرك moving average

تجانس الأسي exponential smoothing

أريما SARIMA

المتوسط المتحرك moving average

ربما يكون نموذج المتوسط المتحرك هو النهج الأكثر سهولة لنمذجة السلاسل الزمنية.

يوضح هذا النموذج ببساطة أن الملاحظة التالية هي متوسط جميع المشاهدات السابقة

على الرغم من بساطته ، قد يكون هذا النموذج جيدًا بشكل مدهش ويمثل نقطة انطلاق جيدة

خلاف ذلك ، يمكن استخدام المتوسط المتحرك لتحديد الاتجاهات المثيرة للاهتمام .في البيانات.

تجانس الأسي exponential smoothing

إن التجانس الأسي مشابهًا للمتوسط المتحرك ، ولكن , يتم تعيين وزن متناقص مختلف لكل ملاحظة. وبعبارة أخرى ، يتم إعطاء أهمية أقل للملاحظات بينما نتحرك أبعد من الحاضر.

رياضيا ، يتم التعبير عن التجانس الأسي على النحو التالي:

$$y = \alpha x_t + (1 - \alpha) y_{t-1}, t > 0$$

alpha هي عامل تجانس يأخذ القيم بين 0 و 1. ويحدد مدى سرعة انخفاض الوزن للملاحظات السابقة

أو يمكننا أستخدام Double exponential smoothing

المزدوج :يتم استخدام التجانس الأسي المزدوج عندما يكون هناك اتجاه في السلسلة الزمنية. وهي ببساطة استخدام للتجانس الأسي مرتين.

رياضيا

$$y = \alpha x_t + (1 - \alpha)(y_{t-1} + b_{t-1})$$
$$b_t = \beta(y_t - y_{t-1}) + (1 - \beta)b_{t-1}$$

أو يمكننا أستخدام ripe exponential smoothing

تعمل هذه الطريقة على توسيع التجانس الأسي المزدوج من خلال إضافة عامل تجانس موسمي. بالطبع ، هذا مفيد إذا لاحظنا وجود الموسمية في السلسلة الزمنية الخاصة بنا.

رياضيا

$$y = \alpha \frac{x_t}{c_{t-L}} + (1 - \alpha)(y_{t-1} + b_{t-1})$$
$$b_t = \beta(y_t - y_{t-1}) + (1 - \beta)b_{t-1}$$
$$c_t = \gamma \frac{x_t}{y_t} + (1 - \gamma)c_{t-L}$$

نموذج المتوسط المتحرك للانحدار الذاتي الموسمي

Seasonal autoregressive integraded moving average model (SARIMA)

هو في الواقع مزيج من نماذج أبسط لصنع نموذج معقد يمكن أن يمثل SARIMA سلسلة زمنية تعرض خصائص غير ثابتة وموسمية.

في البداية , لدينا نموذج الانحدار الذاتي AR (p) هذا في الأساس هو انحدار السلسلة الزمنية على نفسها. هنا , نفترض أن القيمة الحالية تعتمد على قيمها السابقة مع بعض التأخير. حيث يأخذ معلمةp والتي تمثل الحد الأقصى للتأخير.

6-البرمجة التطورية - Evolutionary Programming

تجمع هذه التقنية بين الأنواع المختلفة لتحليل البيانات باستخدام الخوارزميات التطورية. يمكنها استكشاف مساحة بحث واسعة وإدارة تفاعل السمات بشكل فعال للغاية.

7-المنطق الضبابي - Fuzzy Logic

هو تقنية تحليل البيانات على أساس الاحتمالية التي تساعد في التعامل مع الشكوك في تقنيات استخراج البيانات..

8-الشبكات العصبية الاصطناعية - Artificial Neural Networks:

سيشرح بشكل معمق أدناه في قسم التنقيب في المعطيات

منهجيات / طرق التنقيب في المعطيات – Data mining methods

مقدمة:

هناك العديد من الطرق المستخدمة للتنقيب عن البيانات ولكن الخطوة الحاسمة هي تحديد الطريقة المناسبة منها وفقًا للعمل أو بيان المشكلة.

منهجيات / طرق التنقيب في المعطيات :

Association
Classification
Clustering Analysis
Prediction
Sequential Patterns or Pattern Tracking
Decision Trees
Outlier Analysis or Anomaly Analysis
Neural Network

Association :

تشترك هذه المنهجية أيضًا مع علم البيانات لذلك تم شرحها أعلاه

Classification:

التصنيف هي منهجية تقوم بتصنيف البيانات إلى عدد مميز من الفئات وبالتالي يتم تعيين التسمية لكل فئة.

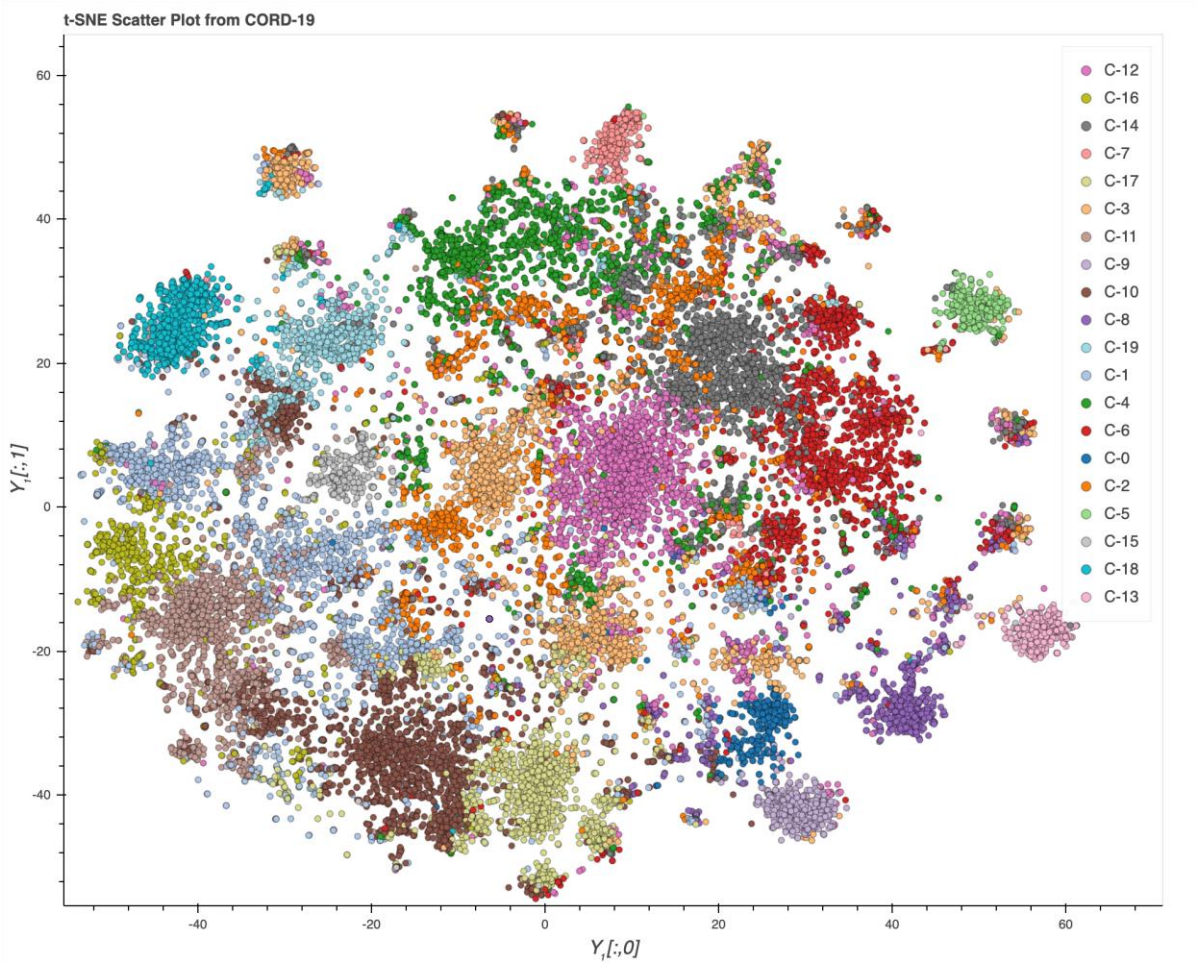
الهدف الرئيسي من التصنيف هو تحديد النمط المخفي من العملية لإطلاق بيانات جديدة من خلال تحليل مجموعة التدريب. بشكل عام ،

كما يمكن إجراء التصنيف على كل من البيانات المنظمة وغير المنظمة.

خوارزميات التصنيف : (سيتم التطرق لهم بالتفصيل في القسم القادم)

Naive Bayes classifier
Decision Trees
Support Vector Machine
Random Forest
K- Nearest Neighbors

: **Clustering Analysis**



منهجية التجميع هي نوع من منهجيات التعلم الآلي مفيدة لفصل مجموعة البيانات بناءً على المجموعات الفردية واحتياجات العمل. تضم فئة شائعة من خوارزمية التعلم الآلي التي يتم تنفيذها في علوم البيانات والذكاء الاصطناعي (AI).

حالات الاستخدام لخوارزميات التجميع هي تجزئة الصورة وتقسيم السوق وتحليل الشبكات الاجتماعية.

أنواع خوارزمية العنقدة:

في الأساس ، تنقسم خوارزمية العنقدة إلى مجموعتين فرعيتين هما:

- **التجميع الصلب:** تنتمي مجموعة من كيانات البيانات المتشابهة إلى سمة أو مجموعة مماثلة تمامًا.

إذا كانت كيانات البيانات ليست مماثلة لشرط معين ، فسيتم إزالة كيان البيانات بالكامل من مجموعة نظام المجموعة.

- 2. **التجميع المرن:** حيث يجد كيان بيانات مشابهًا لتشكيل مجموعة. في هذا النوع من المجموعات ، يمكن العثور على كيان بيانات فريد في مجموعات متعددة تم تعيينها وفقًا لمثلها.

خوارزميات العنقدة: (سيتم التطرق لهم بالتفصيل في القسم القادم)

1. Connectivity Models
2. Centroid Models
3. Distribution Models
4. Density Models
5. K Means Clustering
6. Hierarchical Clustering

تطبيقات خوارزمية التجميع:

يتم استخدامه في الكشف عن الشذوذ
يتم استخدامه في تجزئة الصورة
يتم استخدامه في التصوير الطبي
يتم استخدامه في تجميع نتائج البحث
يتم استخدامه في تحليل الشبكة الاجتماعية
يتم استخدامه في تجزئة السوق
يتم استخدامه في محركات التوصية

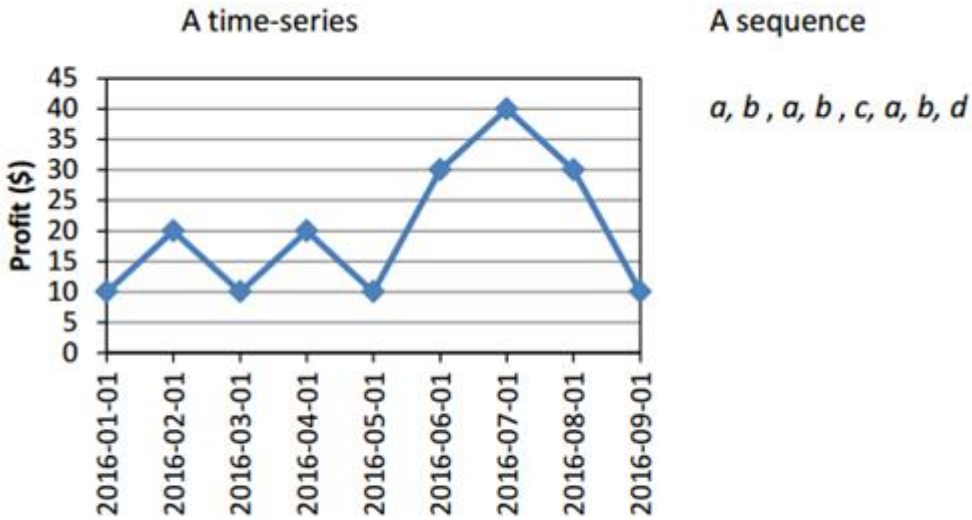
Prediction :

تُستخدم هذه الطريقة للتنبؤ بالمستقبل بناء على الاتجاهات الحالية أو مجموعة البيانات.

يستخدم التنبؤ في الغالب مع مجموعة من طرق التعدين الأخرى مثل التصنيف

فقد تستخدم هذه الطريقة لتوليد نموذج للتوقع مقدار الإيرادات التي سيولدها كل عنصر بناء على بيانات المبيعات السابقة

Sequential patterns or Pattern tracking :

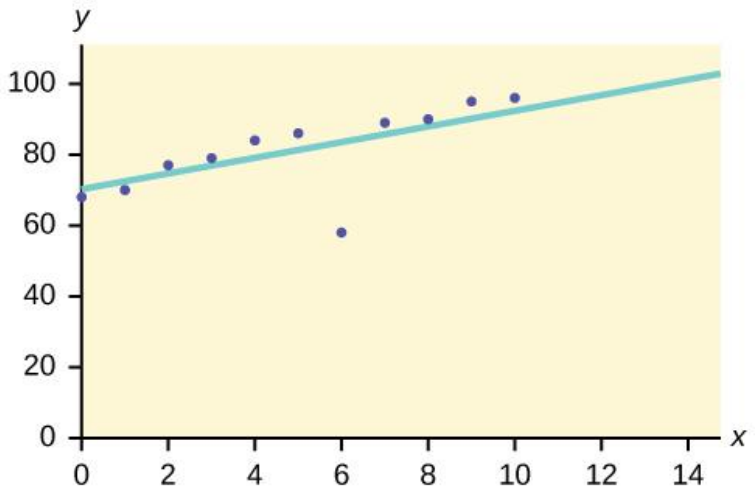


A time-series (left) and a sequence (right)

تُستخدم هذه الطريقة لتحديد الأنماط التي تحدث بشكل متكرر خلال فترة زمنية معينة.

على سبيل المثال ، قد يلاحظ أن الإصابات الخاصة بفيروس كورونا يبدو أنها ترتفع قبل موسم الشتاء مباشرة .

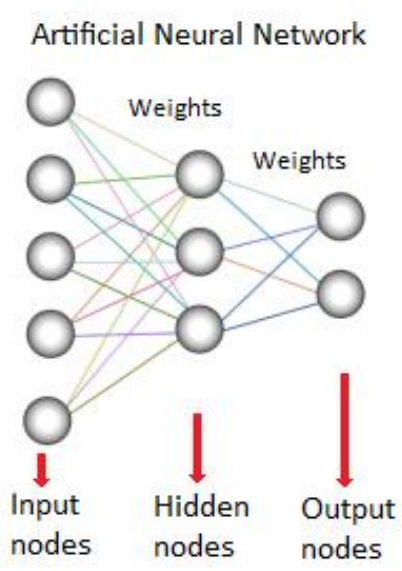
Outlier Analysis or Anomaly Analysis:



يتم استخدام هذه الطريقة لتحديد عناصر البيانات التي لا تتوافق مع النمط المتوقع أو السلوك المتوقع.

حيث تعتبر عناصر البيانات غير المتوقعة هذه شاذة أو ضوضاء. وهي مفيدة في العديد من المجالات مثل الكشف عن الاحتيال في بطاقات الائتمان , وكشف التسلل , واكتشاف الأخطاء وما إلى ذلك. ويسمى هذا أيضًا بالتنقيب الخارجي.

Neural Network:



تعتمد هذه الطريقة أو طريقة النمذجة على الشبكات العصبية البيولوجية. وهي عبارة عن مجموعة من الخلايا العصبية المترابطة مثل وحدات المعالجة ذات الروابط الموزونة(المثقلة) بينها.

يتم استخدامها لنمذجة العلاقة بين المدخلات والمخرجات.

يتم استخدامه لتصنيف وتحليل الانحدار ومعالجة البيانات وما إلى ذلك.

من المعروف أن الشبكات العصبية سهلة الاستخدام للغاية لأنها مؤتمتة إلى حد معين ، ولهذا السبب لا يتوقع أن يكون لدى المستخدم الكثير من المعرفة حول العمل أو قاعدة البيانات. ولكن لكي تعمل الشبكة العصبية بكفاءة ، على المستخدم أن يعرف كيفية توصيل العقد وماهو عدد وحدات المعالجة التي سيتم استخدامها

ومتى يجب إيقاف عملية التدريب.

تمتلك الشبة العصبية أجزاء رئيسية :

- العقدة هي التي تتطابق بحرية مع العصبون في الدماغ البشري
- الارتباط / الوصلات هي التي تتطابق بحرية مع الروابط بين الخلايا العصبية في الدماغ البشري
- بنية الشبكة : وهي تكوين وترابط الخلايا العصبية

Decision Trees:

تم شرح شجرة القرارات بشكل معمق أعلاه في قسم خوارزميات علم البيانات

تعريف:

خوارزميات استخراج البيانات هي فئة خاصة من الخوارزميات المفيدة لتحليل البيانات وتطوير نماذج البيانات لتحديد أنماط ذات معنى. وهي جزء من خوارزميات التعلم الآلي. يتم تنفيذ هذه الخوارزميات من خلال برمجة مختلفة مثل لغة R و Python واستخدام أدوات استخراج البيانات , لاشتقاق نماذج البيانات المحسنة.

C4.5 Algorithm :

تُستخدم خوارزمية C4.5 في التنقيب عن البيانات كمصنف لشجرة القرار والتي يمكن استخدامها لتوليد قرار ، بناءً على عينة معينة من البيانات (تنبؤات أحادية المتغير أو متعددة المتغيرات). حيث أنها تستخدم السمة ذات أعلى اكتساب للمعلومات الطبيعية كمعايير تقسيم.

كما يمكن عند استخدام الخوارزمية العمل مع كل من البيانات المنفصلة والمستمرة

كما يمكنها التعامل مع مشكلة البيانات الغير مكتملة بشكل جيد للغاية.

The k-means Algorithm :

تعد خوارزمية K-Means بأنها خوارزمية تعلم غير خاضعة للإشراف لها عملية تكرارية يتم فيها عنقدة مجموعة البيانات في عدد العناقيد k حيث أنها تجد عدداً ثابتاً (k) من المجموعات في البيانات , يتم تجميعها معاً بسبب أوجه التشابه في ميزاتها

عند استخدام خوارزمية K–Means ، يتم تعريف المجموعة بواسطة محور مركزي ، وهي نقطة (إما خيالية أو حقيقية) في مركز الكتلة. كل نقطة في مجموعة البيانات هي جزء من المجموعة التي تقع النقطة المركزية فيها على مسافة قريبة

أي أنها تجد عدد k من النقط الوسطى ، ثم تقوم بتعيين جميع نقاط البيانات إلى أقرب مجموعة ، بهدف الحفاظ على النقط الوسطى صغيرة.

مما يجعل النقاط الداخلية للمجموعة متشابهة قدر الإمكان أثناء محاولة الحفاظ على المجموعات في مساحة مميزة ،

Naive Bayes Algorithm :

تعتمد هذه الخوارزمية على نظرية بايز.

تستخدم هذه الخوارزمية بشكل رئيسي عندما تكون أبعاد المدخلات عالية. يمكن لهذا المصنف حساب الناتج المحتمل التالي بسهولة. يمكن إضافة بيانات أولية جديدة خلال وقت التشغيل وتوفر تصنيفاً احتمالياً أفضل. تحتوي كل فئة على مجموعة معروفة من المتجهات التي تهدف إلى إنشاء قاعدة تسمح بتعيين الكائنات للفئات في المستقبل. تصف ناقلات المتغيرات الأشياء المستقبلية. وهي واحدة من أسهل الخوارزميات لأنها سهلة البناء ولا تحتوي على أي مخططات معقدة لتقدير المعلمات. كما يمكن تطبيقها بسهولة على مجموعات البيانات الضخمة أيضاً.

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

The Apriori Algorithm:

تستخدم هذه الخوارزمية للعثور على المجموعات المتكررة من مجموعة بيانات المعاملات واستنباط قواعد الارتباط , حيث يتم استخدام خوارزمية Apriori على نطاق واسع.

بمجرد الحصول على مجموعات العناصر المتكررة ، يصبح من الواضح إنشاء قواعد الارتباط للحد الأدنى من الثقة المحددة الأكبر أو المتساوي.

انضمام: يتم استخدام قاعدة البيانات بأكملها لمجموعات مجرفة ا عنصر مجرفة.

التقليم: يجب أن تلبي مجموعة العناصر هذه الدعم والثقة للانتقال إلى الجولة التالية لمجموعتي العناصر.

التكرار: حتى لا يتم الوصول إلى الحجم المحدد مسبقًا حتى يتم تكرار ذلك لكل مستوى مجموعة عناصر.

الفصل الثاني

الإطار النظري

(المبحث الثاني: دراسة الحالة)

"Logic will get you from A to Z; imagination will get you everywhere."

Albert Einstein

المحتوى:

يتضمن الإطار النظري المفاهيم أساسية التي تؤطر عمل الإطار العملي للبحث وهي:

- جائحة فيروس كورونا / **COVID_19**
- مفهوم الجودة
- مفهوم جودة النتائج الطبية

يتناول البحث بشكل أساسي استخدام تقانات تحليل المعطيات لدعم جودة النتائج الطبية , ونقصد بالنتائج الطبية في سياق شرحنا عن الحالة العملية للمصابين بفيروس كورونا \ COVID_19 .

في 31 ديسمبر 2019 تم تنبيه منظمة الصحة العالمية إلى العديد من حالات الالتهاب الرئوي في مدينة ووهان بمقاطعة هوبي في الصين.

لم يتطابق الفيروس مع أي فيروس آخر معروف. الأمر الذي أثار القلق وذلك بسبب تواجد فيروس جديد تكون المعلومات عن كيفية انتشاره وتأثيره على الناس أمر غير مضبوط ومعلوم .

فيروسات كورونا \ Corona Virus:

هي فصيلة فيروسات واسعة الانتشار يُعرف أنها تسبب أمراضاً تتراوح من نزلات البرد الشائعة إلى الأمراض ومتلازمة الالتهاب الرئوي الحاد الوخيم (السارس).

من المتعارف عن فيروسات كورونا انها قد تسبب المرض للحيوان والإنسان , ومن المعروف أن عدداً من فيروسات كورونا تسبب لدى البشر أمراض تنفسية تتراوح حدتها من نزلات البرد الشائعة إلى الأمراض الأشد وخامة مثل متلازمة الشرق الأوسط التنفسية (ميرس) والمتلازمة التنفسية الحادة الوخيمة (سارس). ويسبب فيروس كورونا المُكتشف مؤخراً مرض COVID_19.

سلالة COVID_19:

مرض COVID_19 هو مرض معد يسببه آخر فيروس تم اكتشافه من سلالة فيروسات كورونا.

ولم يكن هناك أي علم بوجود هذه السلالة من الفيروس الجديد ومرضه قبل بدء تفشيه في مدينة ووهان الصينية في كانون الأول / ديسمبر 2019.

وقد تحوّل COVID_19 الآن إلى جائحة تؤثر على العديد من بلدان العالم.

أعراض فيروس COVID_19:

الأعراض الأكثر شيوعاً	الأعراض الأقل شيوعاً
الأعراض التنفسية	الآلام والأوجاع
الحمى	احتقان الأنف
الإرهاق	الصداع
السعال الجاف	التهاب الملتحمة
	ألم الحلق
	الإسهال
	فقدان حاسة الذوق أو الشم
	ظهور طفح جلدي أو تغير لون أصابع اليدين أو القدمين

الشفاء :

وفق ما تتناوله الأحصائيات الخاصة بمنظمة الصحة العالمية ومراكز أبحاث طبية فعليه يتعافى معظم الناس

(نحو 80%) من المرضى دون الحاجة إلى علاج خاص.

ولكن الأعراض تشتد لدى شخص واحد تقريباً من بين كل 5 أشخاص مصابين بمرض **COVID_19** فيعاني من صعوبة في التنفس.

وتزداد مخاطر الإصابة بمضاعفات وخيمة بين المسنين والأشخاص المصابين بمشاكل صحية أخرى مثل ارتفاع ضغط الدم أو أمراض القلب والرئة أو السكري أو السرطان.

آلية إنتشار COVID_19

يمكن أن يلقط الأشخاص عدوى **COVID_19** من أشخاص آخرين مصابين بالفيروس.

وينتشر المرض بشكل أساسي من شخص إلى شخص عن طريق القُطيرات الصغيرة التي يفرزها الشخص المصاب من أنفه أو فمه عندما يسعل أو يعطس أو يتكلم.

وهذه القطيرات وزنها ثقيل نسبياً، فهي لا تنتقل إلى مكان بعيد وإنما تسقط سريعاً على الأرض.

ويمكن أن يلقط الأشخاص مرض **COVID_19** إذا تنفسوا هذه القُطيرات من شخص مصاب بعدوى الفيروس.

وعليه كانت النصائح أن يكون هناك مسافة متر واحد على الأقل (٣ أقدام) من الآخرين.

وقد تحط هذه القطيرات على الأشياء والأسطح المحيطة بالشخص، مثل الطاولات ومقابض الأبواب ودرابزين السلالم. ويمكن حينها أن يصاب الناس بالعدوى عند ملامستهم هذه الأشياء أو الأسطح ثم لمس أعينهم أو أنفهم أو فمهم. لذلك من المهم غسل المواظبة على غسل اليدين بالماء والصابون أو تنظيفهما بمطهر كحولي لفرك اليدين.

الفرق بين العزل الذاتي والحجر الصحي الذاتي والتباعد الجسدي

الحجر الصحي يعني تقييد الأنشطة وعزل الأشخاص غير المرضى هم أنفسهم ولكنهم ربما تعرّضوا للإصابة بعدوى COVID_19. والهدف هو منع انتشار المرض في الوقت الذي لا تكاد تظهر أي أعراض على الشخص.

أما العزل فيعني عزل الأشخاص المرضى الذين تظهر عليهم أعراض COVID_19 ويمكنهم نقل عدواه، لمنع انتشار المرض.

ويعني التباعد الجسدي الابتعاد عن الآخرين جسدياً . وتوصي المنظمة بالابتعاد عن الآخرين مسافة متر واحد (٣ أقدام) على الأقل. وهي توصية عامة يتعين على الجميع تطبيقها حتى لو كانوا بصحة جيدة ولم يتعرضوا لعدوى ..COVID_19

مفهوم الجودة :

شهد مفهوم الجودة في الآونة الأخيرة اهتماما كبيرا من قبل الباحثين الأكاديميين والقادة ومدراء الأعمال وذلك نظرا للتنافس الكبير والتسارع المستمر على التميز في تقديم الخدمات والمنتجات بمستوى عال وذلك لتحسين من إجرائيات العمل أو حتى لتحسين نتائج موجودة من قبل ورفع مستوى دقتها بشكل مستمر على نحو يساعد على الوصول إلى المستوى الأدق في السعي للوصول إلى الأمثلية كحيازة رضى العميل والحفاظ عليه أو تحسين جودة المدخلات الطبية للحصول على مخرج يوصف بدقة نتائجه , وذلك في ظل التنافس الكبير بين الدول ومراكز الأبحاث والشركات وحتى على صعيد الأفراد نتيجة التطور التكنولوجي والصناعي والخدمي الذي يشهده العصر الحالي.

وللجودة تعاريف متعددة يمكن حصرها بالشكل العام بمستويين.

المستوى الأول : القدرة على تقديم المخرجات سواء كانت منتجات أو خدمات بحيث تحقق طموحات العميل فيما يتوقعه من هذا المنتج أو الخدمة

المستوى الثاني : القدرة على تقديم المخرجات سواء كانت منتجات أو خدمات بجيث تحقق معايير قياسية معتمدة من قبل هيئات ناظمة معترف بها.

مفهوم جودة النتائج الطبية

يدرس البحث القائم حالياً استخدام تقانات تحليل المعطيات لدع جودة النتائج الطبية

فعند صياغة جملة جودة النتائج الطبية يقصد بها الباحث استخدام تقانات تحلالت المعطيات وذلك لتحسين ورفع من دقة النتائج الخرجة الطبية عن طريق استخدام منهجيات كاتنبؤ والسلاسل الزمنية , لتقدم للمراكز الطبية ومراكز الأبحاث والمعنيين في مجال الرعاية الصحية رؤية أوضح بالاعتماد على منظور الإحصاء والرياضيات .

حيث تعمل تحليلات البيانات المتطورة إذا تم استخدامها بشكل صحيح ، على تحسين رعاية المرضى في نظام الرعاية الصحية.

فإن تحليل البيانات المتاحة لاكتشاف الممارسات الأكثر فعالية يساعد على خفض التكاليف وتحسين صحة السكان الذين تخدمهم مؤسسات الرعاية الصحية

تشير "تحليلات البيانات" إلى ممارسة أخذ كميات كبيرة من البيانات المجمعة وتحليلها من أجل استخلاص رؤى ومعلومات مهمة واردة فيها. هذه العملية مدعومة بشكل متزايد بالبرامج والتكنولوجيا الجديدة التي تساعد على فحص كميات كبيرة من البيانات بحثاً عن المعلومات المخفية.

في سياق نظام الرعاية الصحية ، الذي يعتمد بشكل متزايد على البيانات ، يمكن لتحليلات البيانات أن تساعد في استخلاص رؤى قد تكون مكملة ومساعدة للنتائج الطبية الحالية،

حيث يمكن تتبع الحالة الصحية للفرد أو للمجتمع ويمكن حتى تتبع وتحديد الأشخاص المعرضين لخطر الإصابة بفيروس كورونا . من خلال هذه المعلومات الإضافية ، يمكن

للنظام الصحي تخصيص الموارد بشكل أكثر كفاءة أو قد تقدم له نظرة على موضوع قد كان مغيب وذلك من أجل زيادة دقة النتائج التي بدورها تصب في صحة الفرد وصحة السكان ، والأهم من ذلك ، رعاية المرضى.

توقع المخاطر:

ينطوي علاج الأمراض المزمنة على واحدة من أكبر تكاليف صناعة الرعاية الصحية.

على المستوى السكاني ، يمكن للتحليلات التنبؤية أن تساعد بشكل كبير في خفض التكاليف من خلال التنبؤ بالمرضى الأكثر عرضة للإصابة بالأمراض وترتيب التدخل المبكر ، قبل ظهور المشاكل. أو من خلال المساعدة في التنبؤ بعد من الإصابات المتوقع حدوثها بفترات لاحقة, يتضمن هذا تجميع البيانات المتعلقة بمجموعة متنوعة من العوامل. وتشمل هذه التاريخ الطبي ، والوضع الديموغرافي والاجتماعي والاقتصادي والأمراض المصاحبة.

يتضمن التاريخ الطبي عادة العمر وضغط الدم والجلوكوز في الدم والتاريخ العائلي للحالات المزمنة ومستويات الكوليسترول , بالإضافة إلى باقي مجموعة الأمراض المزمنة المصاحبة للمرض.

ترتبط نسبة كبيرة مما يؤثر على جودة النتائج الصحية بعوامل خارج نطاق الرعاية الصحية التقليدية. تشمل هذه العوامل العادات والسلوكيات الصحية للمرضى ، والعوامل الاجتماعية الاقتصادية مثل التوظيف والتعليم ، والبيئة المادية. من أجل تحسين النتائج ، يجب على نظام الصحة العامة توسيع حدوده لمراعاة هذه العوامل

"الخارجية". في تحليلات البيانات ، يمكن نمذجة هذه المقاييس للتنبؤ بخطر الإصابة بالأمراض المزمنة.

أخيرًا ، يجب أن تشكل التحليلات نموذجًا للمخاطر من خلال احتساب الحالات الطبية المتعددة التي قد يعاني منها المريض. عند تجميع وتحليل كل هذه الأشكال من البيانات ، يمكن لصناعة الرعاية الصحية تخصيص الموارد بشكل أكثر فعالية ، مما يمكنها من التدخل بقوة في السكان المعرضين لخطر كبير في وقت مبكر ومنع التكاليف النظامية طويلة الأجل

الفصل الثالث

الإطار العملي للبحث

"The world as we have created it is a process of our thinking. It cannot be changed without changing our thinking."

Albert Einstein

المحتوى:

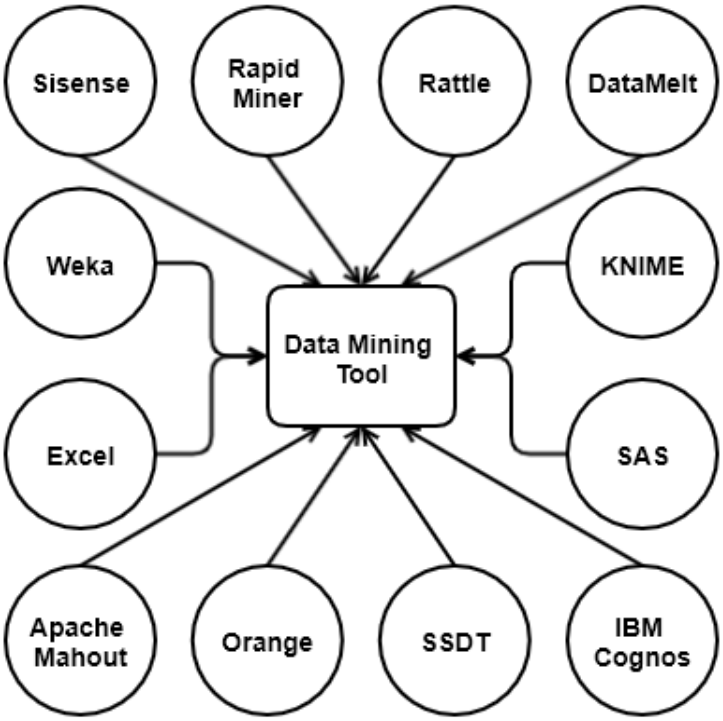
يتضمن الفصل الثالث الإطار العملي للبحث والذي يحوي على التالي:

- الأدوات المستخدمة في تحليل البيانات والتنقيب عن المعطيات
- شرح مجموعة البيانات
- تنظيف البيانات
- المقاربة الأولى: إقتراح نموذج للتنبؤ بعدد الإصابات والوفيات وحالات الشفاء المتوقع حدوثها خلال يوم معين بناء على البيانات السابقة
- المقاربة الثانية: إقتراح نموذج للتنبؤ بإتجاه إنتشار المرض والتنبؤ بعدد الإصابات التي ستع بفترات مستقبلية.
- المقاربة الثالثة: إقتراح نموذج لملاحظة نسبة نمو الفيروس في مجموعة من البلدان وذلك لوضع مقياس للبلدان التي قد وصل عدد الإصابات فيها إلى الذروة
- المقاربة الرابعة: إقتراح نموذج لتحسين جودة النتائج الطبية انطلاقا من تجميع بيانات أعراض الإصابات وفق عناقيد لاكتشاف الأنماط المخفية

الأدوات المستخدمة في تحليل البيانات والتنقيب عن المعطيات:

تعريف

في عالم اليوم , يتم إنشاء كمية كبيرة من البيانات في غضون ثوان. حيث أنه من الضروري عند التعامل مع هذه البيانات , أن يكون لدينا بعض المعرفة بالتقنيات والأدوات المختلفة. أدوات التنقيب وتحليل البيانات ليست سوى مجموعة من المنهجيات المستخدمة لتحليل هذه الكمية الكبيرة من البيانات والعلاقة بين البيانات المختلفة.



الأدوات المستخدمة في هذا البحث :

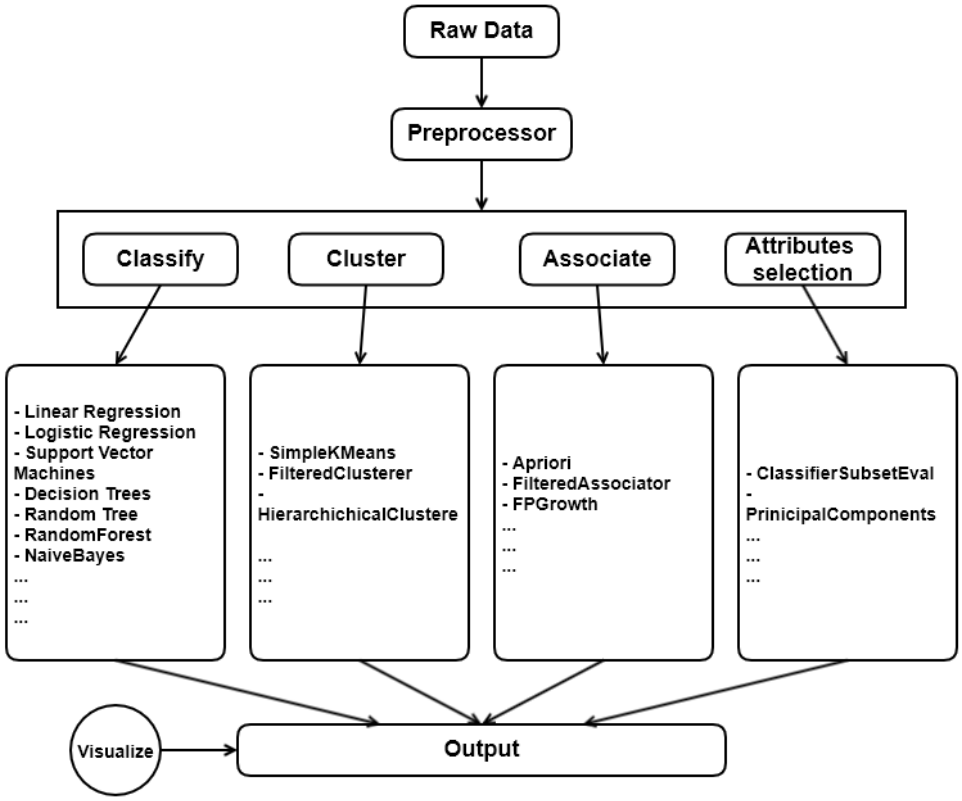
Excel

Excel هو تطبيق جداول بيانات تم تطويره ونشره بواسطة شركة Microsoft وهو جزء من مجموعة برامج Microsoft Office الخاصة بالبرامج الإنتاجية.

هو برنامج مفيد وفعال لتحليل البيانات وتوثيقها. حيث أن البرنامج عبارة عن جدول بيانات يحتوي على عدد من الأعمدة والصفوف , حيث يكون كل تقاطع بين العمود والصف "خلية". تحتوي كل خلية على نقطة واحدة من البيانات أو معلومة واحدة. من خلال تنظيم المعلومات بهذه الطريقة , يمكننا تسهيل العثور على المعلومات ورسم المعلومات تلقائيًا من تغيير البيانات

Weka

يعد برنامج Weka برنامج شامل يتيح لنا معالجة البيانات الكبيرة مسبقًا وتطبيق خوارزميات تعلم الآلة المختلفة على البيانات الكبيرة ومقارنة المخرجات المختلفة. يعمل هذا البرنامج على تسهيل العمل مع البيانات الضخمة وتدريب الآلة باستخدام خوارزميات التعلم الآلي وقدرته على تصوير البيانات (عرضها بشكل مرئي)



Orange:

عبارة عن مجموعة أدوات لتصور البيانات مفتوحة المصدر والتعلم الآلي واستخراج البيانات. يتميز بواجهة برمجة بصرية أمامية لتحليل البيانات الاستكشافية والتصور التفاعلي للبيانات, كما يتم استخدام برنامج orange في التعلم الآلي ، والتنقيب عن البيانات ، وتحليل البيانات.

الأدوات المستخدمة في تحليل وتنقيب البيانات وسبب استخدامها

Excel

استخدم الباحث أداة تحليل البيانات Excel في تحليل البيانات الخاصة بالإصابات الناجمة عن مرض فيروس كورونا وذلك لقدرتها التحليلية في استثمار الأعمدة والصفوف الخاصة به من خلال إجراء التالي :

time seies

linear regression

logistic growth model

كما تم إجراء تصوير لتلك النماذج لتسهل على الباحث والمهتمين في البحث قرائة النتائج

Weka

تم الإستعانة ببرنامج Weka مفتوح المصدر الذي يعتبر البديل الأسهل في إجراء وتطبيق خوارزميات التنقيب في المعطيات و التصور مرئي لها , حيث تم استخدام برنامج Weka في تطبيق خوارزمية k means على البيانات الخاصة بالأعراض الناجمة عن العدوى من فيروس كورونا , ذلك لقدرته على إجراء العديد من خوارزميات العنقدة بسرعة وسهولة عالية مقارنة بمنافسيها من البرمجيات

كما يتميز برنامج Wekaبمكتبته الواسعة من الخوارزميات دون الحاجة إلى ضرورة الإستعانة بمكاتب برمجية وإدخال أي اكواد لغات برمجية مختلفة

كما استعان الباحث بخاصية تنظيف البيانات الموجودة ضمن البرنامج في الواجهة الرئيسية من خيار edit وذلك للمقارنة بالنتائج المنظفة الخاصة بها مع النتائج المنظفة المخرجة من برنامج excel

أما بالنسبة لصيغة مجموعة البيانات المحملة من قبل الباحث هي .CSV والتي يسهل قرائتها وجلبها ومعالجتها بسهولة وسرعة من قبل برنامج ويكا

Orange

استعان الباحث بمميزات برنامج Orange وذلك لقدرته على إحتساب عدد العناقيد k الأفضل الواجب إضافتها والتوزيع عليها ضمن عملية تنقيب النمعطيات باستخدام خوارزمية K means

ملاحظة:

بعد أن حدد البرنامج k=2 أجرينا نظرة موسعة على البيانات ليتبين لنا أن استخدام عنقودين من شأنه أن يبين لنا نتائج جيدة ولكن من الممكن أن نحصل على نتائج أفضل إذا تم استخدام 8 عناقيد بناء على الملاحظة والتدقيق .

شرح مجموعة البيانات

١- مجموعة بيانات Novel Corona Virus 2019

وهي معلومات على مستوى اليوم عن الحالات المصابة بفيروس كورونا

مصدر البيانات :المصدر johns Hopkins university

منذ 22 يناير 2020.

تم تجميع البيانات من قبل مركز جامعة جونز هوبكنز لعلوم وهندسة النظم

(JHU CCSE) من مصادر مختلفة بما في ذلك منظمة الصحة العالمية , لجنة الصحة الوطنية لجمهورية الصين الشعبية , وزارة الصحة في هونغ كونغ , حكومة كندا , وزارة الصحة الحكومية الأسترالية , ، وزارة الصحة في سنغافورة

المحتوى:

تحوي مجموعة البيانات على التسلسل الزمني للإصابات اليومية بفيروس كورونا في أنحاء العالم , وعليه تشمل الحقول المتاحة في البيانات

المقاطعة / الولاية ، البلد / المنطقة / آخر تحديث للبيانات / الإصابات / الحالات مشتبه به / حالات الشفاء / عدد الوفيات / كما تحوي درجات الطول والعرض لتموضع الإصابة

الفترة الزمنية:

يبدأ تسجيل الإصابات في مجموعة البيانات بتاريخ 1/22/2020 (الخلية الأولى) لتنتهي في تاريخ 6/30/2020 (الخلية الأخيرة) .. إن البيانات تحدث بشكل يومى .. إلا أن الفترة التي تم فيها تحميل البيانات كانت في نهاية الشهر السادس

أبعاد حجم مجموعة البيانات:

كما تحوي : عدد الصفوف 64133 / عدد الأعمدة 8

ملاحظة : هذه بيانات سلسلة زمنية , أي عدد الحالات في أي يوم هو العدد التراكمي

البلدان :

تحوي مجموعة البيانات على ١٩٤ بلد حول العالم تم إجراء الدراسة عليهم

Botswana	Zimbabwe	Suriname	Portugal	Mongolia	Kazakhstan	Germany	Cuba	Australia	Afghanistan
Burundi	Canada	Sweden	Qatar	Montenegro	Kenya	Ghana	Cyprus	Austria	Albania
Sierra Leone	Dominica	Switzerland	Romania	Morocco	Korea, South	Greece	Czechia	Azerbaijan	Algeria
Netherlands	Grenada	Taiwan*	Russia	Namibia	Kuwait	Guatemala	Denmark	Bahamas	Andorra
Malawi	Mozambique	Tanzania	Rwanda	Nepal	Kyrgyzstan	Guinea	Djibouti	Bahrain	Angola
United Kingdom	Syria	Thailand	Saint Lucia	Netherlands	Latvia	Haiti	Dominican Republic	Bangladesh	Antigua and Barbuda
France	Timor-Leste	Togo	Saint Vincent and the Grenadines	New Zealand	Lebanon	Holy See	Ecuador	Barbados	Argentina
South Sudan	Belize	Trinidad and Tobago	San Marino	Nicaragua	Liberia	Honduras	Egypt	Belarus	Armenia
Western Sahara	Laos	Tunisia	Saudi Arabia	Niger	Liechtenstein	Hungary	El Salvador	Canada	Belgium
Sao Tome and Principe	Libya	Turkey	Senegal	Nigeria	Lithuania	Iceland	Equatorial Guinea	Central African Republic	Benin
Yemen	West Bank and Gaza	Uganda	Serbia	North Macedonia	Luxembourg	India	Eritrea	Chad	Bhutan
	Guinea-Bissau	Ukraine	Seychelles	Norway	Madagascar	Indonesia	Estonia	Chile	Bolivia
	Mali	United Arab Emirates	Singapore	Oman	Malaysia	Iran	Eswatini	China	Bosnia and Herzegovina
	Saint Kitts and Nevis	United Kingdom	Slovakia	Pakistan	Maldives	Iraq	Ethiopia	Colombia	Brazil
	Canada	Uruguay	Slovenia	Panama	Malta	Ireland	Fiji	Congo (Brazzaville)	Brunei
	Canada	US	Somalia	Papua New Guinea	Mauritania	Israel	Finland	Congo (Kinshasa)	Bulgaria
	Kosovo	Uzbekistan	South Africa	Paraguay	Mauritius	Italy	France	Costa Rica	Burkina Faso
	Comoros	Venezuela	Spain	Peru	Mexico	Jamaica	Gabon	Cote d'Ivoire	Cabo Verde
	Tajikistan	Vietnam	Sri Lanka	Philippines	Moldova	Japan	Gambia	Croatia	Cambodia
	Lesotho	Botswana	Sudan	Poland	Monaco	Jordan	Georgia	Diamond Princess	Cameroon

2- مجموعة بيانات WHO

مجموعة بيانات الأعراض المصاحبة لفيروس كورونا وفق المبادئ التوجيهية التي قدمتها منظمة الصحة العالمية (WHO) .

تحتوي مجموعة البيانات على سبعة متغيرات رئيسية سيكون لها تأثير على ما إذا كان شخص ما مصابًا بمرض فيروس تاجي أم لا ، وصف كل سمة على النحو التالي :

البلد: قائمة الدول التي زارها الشخص أو أقام بها

العمر: تصنيف الفئة العمرية لكل شخص ، وفقًا لمعيار الفئة العمرية لمنظمة الصحة العالمية

0 to 9	10 to19	20 to24	25 to59	60+

الأعراض: وفقًا لمنظمة الصحة العالمية ، (5) أعراض رئيسية:

الحمى	التعب	صعوبة التنفس	السعال الجاف	التهاب الحلق

كما يوجد أعراض أخرى : الآلام واحتقان الأنف وسيلان الأنف والإسهال وغيرها.

الخطورة: مستوى الشدة ، خفيف ، معتدل ، شديد

جهة الاتصال: هل اتصل الشخص مباشرة بمريض COVID-19 آخر

أبعاد حجم مجموعة البيانات:

تحوي : عدد الصفوف 140801 / عدد الأعمدة 16

محتوى الأسطر:

Country	Contact	Severity	Gender
Age	None_Experiencing	Diarrhea	Runny-Nose
Nasal-Congestion	Pains	None_Sympton	Sore-Throat
Difficulty-in-Breathing	Dry-Cough	Tiredness	Fever

القيم التي تأخذها السمات:

Yes or No	Fever
Yes or No	Tiredness
Yes or No	Dry-Cough
Yes or No	Difficulty-in-Breathing
Yes or No	Sore-Throat
Yes or No	None_Sympton
Yes or No	Pains
Yes or No	Nasal-Congestion
Yes or No	Runny-Nose
Yes or No	Diarrhea
Yes or No	None_Experiencing
0 to 9 / 10 th 19 / 20 to 24 / 25 to 59 / 60+	Age
Male / Female	Gender
Mild / Moderate / Severe	Severity
Contact_Yes Contact_No	Contact
China Italy Iran Republic of Korean France Spain Germany UAE Other	Country

تنظيف البيانات

اجرى الباحث مجموعة من الإختبارات الخاصة بتنظيف البيانات وذلك لضمان اتساق البيانات وتوفير البيانات ذات الصلة للمرحلة اللاحقة في التطبيق العملي

I – Accuracy:

تم تحليل البيانات الخاصة بالتسلسل الزمني الخاص بعدد الإصابات (بيانات جامعة جونز هوبكينز) والبيانات الخاصة بمنظمة الصحة العالمية (WHO) وذلك لضمان الدرجة التي تكون فيها البيانات قريبة من القيم الحقيقية.

حيث انه قد قام الباحث بتحديد جميع القيم الصالحة الممكنة بالكشف عن القيم غير الصالحة بسهولة , وذلك بالتنقيح في البيانات المسجلة ومقارنتها مع ما هو موجود على أرض الواقع عبر البحث والتدقيق عليها عبر الانترنت , وذلك للتأكد من صحة وجود المناطق التي ذكر تواجد الإصابات بها .. كما تم مراجعة ماهية الأعراض الملازمة للإصابات وفق الكتيب المنشور من قبل منظمة الصحة , وتم مقارنة عدد الإصابات المسجلة في البلدان مع تلك الموجودة لدينا في مجموعة البيانات عبر منصة تتبع الأوبئة الخاصة لفوغل وذلك لمقارنة البيانات الفعلية الموجودة بالبيانات التي يتم تدقيقها.

وخلصت نتيجة الإختبار الأول بأن البيانات الفعلية تساوي البيانات المدخلة التي يعمل عليها الباحث.

2- Completeness

تم تحميل البيانات من موقع Kaggle الخاص بنشر مجموعات البيانات التابعة للمنظمات والجامعات والشركات .

ومن أجل التحقق من صحة البيانات المحملة .. قام الباحث بتتبع المسار للبيانات لموقع جامعة جونزهوبكينز وموقع منظمة الصحة العالمية ومقارنة النتائج التي تم رفعها على الموقع Kaggle مع تلك التي تم نشرها من قبل المصدر .

وخلصت نتيجة الإختبار الثاني إلى أن البيانات بين المصدرين كانت متطابقة.

3- Consistency

درجة اتساق البيانات داخل نفس مجموعة البيانات أو عبر مجموعات بيانات متعددة.

في الملفات التي تم تحميلها .. كانت تحتوي على العديد من جداول البيانات كل منها منسق بطريقة ليتم العمل على كل واحد منها بأدوات مختلفة .. فتم التحقق من درجة الاتساق للبيانات داخل مجموعات بيانات متعددة لنتأكد من خلوها من التعارض.

وخلصت النتيجة إلى وجود خطأ إملائي وحيد في إحدى خلايا مجموعات البيانات , ليتم تصحيح هذا الخطأ عبر نسخ الخلية الصحيحة من المصدر الأساسي لها ونقلها إلى مجموعة البيانات التي يتم العمل عليها .

4- Uniformity

قام الباحث بتدقيق التنسيقات المتواجدة في مجموعة البيانات ليتحقق من صحة التنسيق وتعميمه على جميع البيانات ذات نفس النوع

تم التدقيق في التنسيق التاريخي الخاص بالبيانات التاريخية ليكون من الشكل (X/XX/XXXX)

كمثال 1/22/2020

وخلصت النتائج إلى ان كامل البيانات التاريخية منسقة بالشكل الصحيح الذي سيمكن الباحث في الإطار العملي من استخدامها للتنبؤ بالإصابات باستخدام نموذج السلاسل الزمنية.

5- Inspection

فحص البيانات للكشف عن الأخطاء عن طريق :

- **Data profiling**

حيث تم التحقق مما إذا كان عمود معين يتوافق مع معايير الأعمدة الباقية كما تم التحقق من كونه (عمود) مسجل كسلسلة أو رقم ؟. بما يتوافق مع البيانات كاملة .

6- Cleaning

يتضمن تنظيف البيانات تقنيات مختلفة بناء على المشكلة ونوع البيانات. يمكن تطبيق طرق مختلفة لكل منها مقايضتها الخاصة.

بشكل عام ، تتم إزالة البيانات غير الصحيحة أو تصحيحها أو حسابها.

- **relevant data**

عالجنا في عملية تنظيف البيانات مجموعة البيانات التي كانت غير ذات صلة وغير مطلوبة تحليلها للنتائج المراد الحصول عليها ولا تتناسب مع سياق المشكلة التي نحاول حلها

في مجموعة البيانات التابعة لجامعة جونز هوبكينز كانت تحوي على السمة (Last Update) .. تم حذف هذه السمة لكونها لا تشكل قيمة مضافة إلى مجموعة البيانات

كما تم حذف السمة الخاصة بالموقع الجغرافي للإصابة كخطوط طول وعرض , وقد تم معالجة السمة Gender التي كان فضاء احتمالاتها داخل البيانات

(Gender_Female \ Gender_Male \ Gender_Transgender) حيث تم حذف السمة Gender_Transgender من كافة مجموعة البيانات وذلك لكونها لا ترتبط بطبيعة النتائج المراد الحصول عليها .

- **Duplicates**

لوحظ أثناء عملية التنظيف أن بعض الملفات الموجودة في المجلد المحمل من جامعة جونز هوبكينز قد كان يحوي ملفات تتشابه بها البيانات أي كان يوجد تكرار لنفس المعلومة

خلصت النتيجة إلى الإستعانة بمصدر واحد من المصادر لكونه يحمل بعض القيم المكررة في باقي المصادر .

- **Type conversion**

تم التأكد من صحة القيم المدخلة داخل البيانات ,على انها تتبع النوع المخزن الصحيح أي ان السمات التي تحتاج إلى بيانات رقمية يجب تكون كلها رقمية .. وفي حال

البيانات التي تخزن ككائن تاريخ يجب أن يخزن كائن تاريخي وفي حال الحاجة يتم التحويل القيم الفئوية من وإلى الأرقام.

خلصت النتيجة إلى أن البيانات الخاصة بالتسلسل الزمني للإصابات سليمة ولا تحتاج إلى تبديل , أما بالنسبة للبيانات الخاصة بالأعراض الناجمة عن الإصابات فكانت بعض السمات يجب أن يتم تبديل القيم داخلها من (1 and 0) بما يتناسب مع المدخلات الواجبة للبرنامج فحولت إلى (Yes and No) وذلك للقيام بمحاولة قرائتها على برنامج Weka لتسهيل عملية تفسير النتائج .

- **Syntax errors**

تم إجراء بحث كامل على البيانات الموجودة في مجموعة البيانات للتأكد من خلوها من أخطاء بناء الجمل كوجود فراغات بين الجمل أو الكلمات , على سبيل المثال ، فاصلة منقوطة مفقودة في نهاية السطر أو قوس إضافي في نهاية دالة قد ينتج عنها خطأ في بناء الجملة

خلصت النتيجة إلى عدم وجود أي من أخطاء بناء الجملة , كما تم التأكد من صحة الدالات المدخلة على برنامج Excel عند تطبيق دالة دمج بيانات 3 أعمدة بعامود واحد ليسهل عملية معالجته على البرنامج () وخلوها من أي فواصل أو أقواس زائدة .

- **Pad strings**

تم التأكد من أن القيم الموجودة في قاعدة البيانات ليست مبطنة بأحرف بادئة أو لاحقة بطول إجمالي محدد. حيث من الممكن أن يمكن أن حرف الحشو مسافة أو حرفًا محددًا.

- **Fix typos**:

تم التأكد من أن القيم الموجودة في قاعدة البيانات الأعراض خالية من الأخطاء المطبعية .

وتم مراجعة البيانات الموجودة في مجموعة بيانات السلسلة الزمنية والتدقيق في أسماء البلدان

ووصلنا إلى أن مجموعة مؤلفة من 5 بلدان قد تم إستبدال الأحرف الكبيرة بأحرف صغيرة وهذا من شأنه أن يخلق فروقات في مرحلة تنقيب البيانات ليتم حساب كل منها بلد مغايراً عن البلد الأصلي

تم استخدام تطابق النمط لتصحيح الفروقات.

- **Missing values**

تم التدقيق في مجموعات البيانات للتأكد من خلولها من قيم المفقودة وتم ملاحظة التالي :

وبالنظر إلى حقيقة أن القيم المفقودة لا يمكن تجنبها فعمدنا على حذف القيم المفقودة الموجودة داخل بيانات الأعراض في السمة Contact وذلك لتكون عملية التنقيب أدق.

- **Verifying**

بعد القيام بجميع الأنشطة السابقة الخاصة بتنظيف البيانات تم التحقق في النهاية من صحة البيانات عن طريق إعادة فحصها والتأكد من أن السمات وعدد الصفوف قائم وفق ما تم تعديله لا أكثر من ذلك .

ملاحظة :

عند القيام بتحليل سلاسل زمنية على مجموعة بيانات جونز هوبكينز وجب اتباع تنظيف للبيانات خلال مراحل متقدمة من مرحلة التنظيف الأساسية
فعمدنا على دمج السمة الخاصة في العمر بعامود واحد .. حيث أن الأخيرة كانت موزعة على 5 أعمدة مما يستصعب على برنامج التنقيب في المعطيات معالجتها بذاك الشكل .

المقاربة الأولى: إقتراح نموذج للتنبؤ بعدد الإصابات والوفيات وحالات الشفاء المتوقع حدوثها خلال يوم معين بناء على البيانات السابقة

التطبيق العملي باستخدام : Linear regression analysis

سنستخدام تحليل الإنحدار وذلك لتقدير العلاقات بين متغير تابع واخر مستق ولكي يعطي التحليل قيمة متوقعة للمعيار الناتج عن مجموعة خطية من المتنبئات.

حيث سيتم التطبيق على بيانات

johns hopkins university بيانات الإصابات – الأموات – المتعافيين

على مجموعة البلدان التالية : **South Korea – Syria – Italia**

ملاحظة :البيانات الموجودة في مجموعة البيانات مجمعة تراكميا أي أن القيمة التي يتم التنبؤ بها هي قيمة تراكمية يمكننا طرحها من سابقتها لمعرفة العدد الفقلي في ذلك التاريخ ..(سنناقش القيم الفعلية وليست التراكمية في القسم العملي للسلاسل الزمنية)

Italia:

1– الإصابات والأموات

المتغير التابع المراد التنبؤ به : الأموات (عدد الأموات)

المتغيرات المستقلة الذي يؤثر على المتغير التابع : الإصابات (عدد الإصابات)

بعد إجراء التحليل خلصت النتائج إلى التالي :

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.997303362
R Square	0.994613996
Adjusted R Square	0.994582682
Standard Error	1085.43938
Observations	174

ANOVA

	df	SS	MS	F	Significance F
Regression	1	37422044684	37422044684	31762.62338	4.7167E-197
Residual	172	202646727.5	1178178.649		
Total	173	37624691411			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	-911.5051925	136.4531671	-6.67998561	3.18029E-10	-1180.843572	-642.167	-1180.843572	-642.1668127
X الإصابات	0.144756582	0.000812232	178.2207153	4.7167E-197	0.143153356	0.14636	0.143153356	0.146359808

التفسير:

Multiple R:

إن قوة العلاقة الخطية بين متغيرين هي **0.997303362** اي انها قريبة من الواحد (1)

ومنه نستنتج أن العلاقة الخطية بين المتغيرين علاقة قوية وقد تكون قوية جدا

R Square:

معامل التحديد في مثالنا السابق **0.994613996**

في مثالنا , R2 هو **0.99** ، وهو أمر جيد جداً . هذا يعني أن **99**٪ من قيمنا تناسب نموذج تحليل الانحدار. بمعنى آخر ، يتم تفسير **99**٪ من المتغيرات التابعة (قيم y) بواسطة المتغيرات المستقلة (قيم x). بشكل عام ،

Observations:

174= وهو عدد المشاهدات المدخلة في النموذج

الجزء الثاني من الناتج هو تحليل التباين (ANOVA): نادراً ما يتم استخدام جزء ANOVA لتحليل انحدار خطي بسيط في Excel

ولكن لا باس في الإطلاع على النتائج:

df هو عدد درجات الحرية المرتبطة بمصادر التباين.

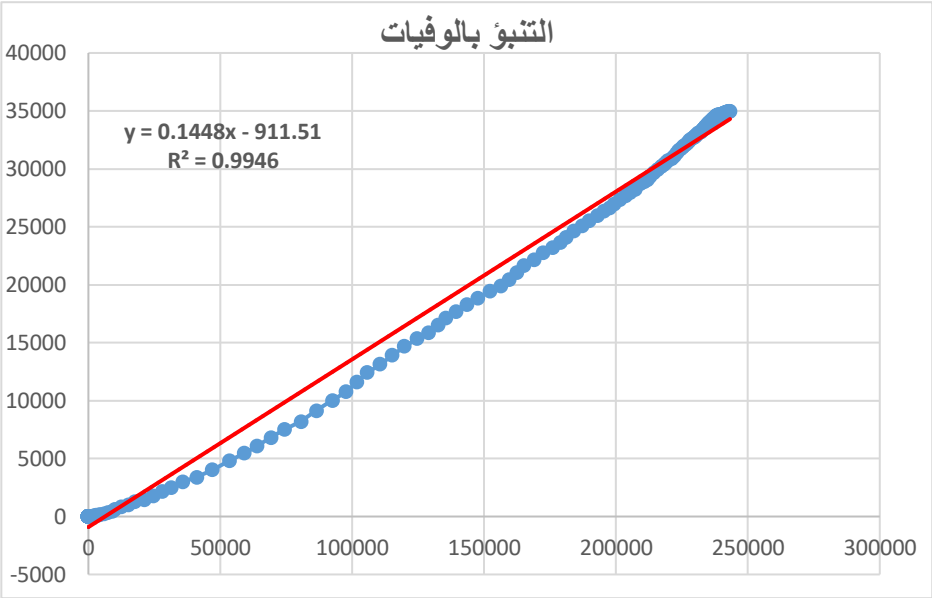
SS هو مجموع المربعات. كلما صغر حجم SS المتبقي مقارنة بـ Total SS ، كان نموذجنا يناسب البيانات بشكل أفضل.

وعليه نلاحظ انا ss هو اصغر من ss Total أي ان النموذج يتناسب مع البيانات

MS هو مربع الوسط

Significance F الخاصة بالنموذج هي أقل من **0.05(5%)** ومنه نستنتج اننا لسنا في حاجة لتغيير المتغير المستقل ,ففي حال كانت Significance F اكبر من 0.05فيجب تغيير المتغير المستقل

النموذج البياني :



يتم صياغة معادلة الانحدار الخطي رياضيا: $y = bx + a + \varepsilon$

حيث أن

- **x** متغير مستقل (الأموات)
- **y** متغير تابع (الإصابات)

ملاحظة:

a هي تقاطع Y ، وهي القيمة المتوسطة المتوقعة لـ **y** عندما تكون جميع المتغيرات **x** تساوي **0**. في الرسم البياني للانحدار ، هي النقطة التي يتقاطع فيها الخط مع المحور Y.

b هو ميل خط الانحدار ، وهو معدل التغيير لـ **y** مع تغير **x**.

ε هو مصطلح الخطأ العشوائي ، وهو الفرق بين القيمة الفعلية لمتغير تابع وقيمته المتوقعة.

يتم حساب الخطأ المعيار تلقائيا في برنامج اكسل

لتكون المعادلة بالشكل النهائي : $y = bx + a$

$$Injuries = b * The\ dead + a$$

بالتعويض في معادية المستقيم : $y = 0.1448x - 911.51$

169	7/7/2020	241956	34899
170	7/8/2020	242149	34914
171	7/9/2020	242363	34926
172	7/10/2020	242639	34938
173	7/11/2020	242827	34945
174	7/12/2020	243061	34954
175	7/13/2020	243230	34967
176	7/14/2020	243344	34324.7
177			
178	0.1448x - 911.51		
179	R^2 = 0.9946		
180			

وعليه نعوض المعادلة المستخرجة من تحليل الانحدار بقيمة البيانات الخاصة بعدد الإصابات بتاريخ 7/14/2020 حتى نتنبأ بعدد الأموات المتوقع وقوعها في هذا اليوم

$y = 0.1448 * 243344 - 911.51$

= 34324 وعليه يمكننا معرفة عدد الأشخاص الذي من المتوقع موتهم جراء الإصابة بالفيروس في تاريخ 7/14/2020 , أما بالنسبة للعدد الفعلي الذي حدث في ذاك اليوم فكان =34984 وهو رقم قريب من القيمة المتوقعة .

والسبب في وجود فارق ولو بسيط بين القيمة المتوقعة والقيمة الحقيقية هو ان تحليل الإنحدار يعطي قيمة متوقعة أي ليست حقيقية للمعيار الناتج عن مجموعة خطية من المتنبئات.

كما هو مبين في الرسم البياني فإن العلاقة إيجابية: إذ أظهر الخط اتجاها ً تصاعديا ً .

هذا يشير إلى أنه مع زيادة المتغير المستقل ، يزداد المتغير التابع أيض ًا.

2– الإصابات والتعافي

المتغير التابع المراد التنبؤ به : المتعافيين (عدد المتعافيين)

المتغيرات المستقلة الذي يؤثر على المتغير التابع : الإصابات (عدد الإصابات)

بعد إجراء التحليل خلصت النتائج إلى التالي :

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.902077858
R Square	0.813744463
Adjusted R Square	0.812661582
Standard Error	32664.04179
Observations	174

ANOVA

	Df	SS	MS	F	Significance F
Regression	1	8.01765E+11	8.01765E+11	751.4624773	1.14057E-64
Residual	172	1.83514E+11	1066939626		
Total	173	9.85279E+11			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	-15584.88998	4106.27441	-3.795384434	0.000204036	-23690.06863	-7479.711323	-23690.06863	-7479.711323
X الإصابات	0.670035997	0.024442436	27.41281593	1.14057E-64	0.621790242	0.718281752	0.621790242	0.718281752

التفسير:

Multiple R:

إن قوة العلاقة الخطية بين متغيرين هي **0.902077858** اي انها قريبة من الواحد (1) ومنه نستنتج أن العلاقة الخطية بين المتغيرين علاقة قوية (ولكنها ليست قوية جداً)

R Square:

معامل التحديد في مثالنا السابق 0.813744463

في مثالنا ، R2 هو 0.81 ، وهو أمر جيد نوعاً ما. هذا يعني أن 81٪ من قيمنا تناسب نموذج تحليل الانحدار. بمعنى آخر ، يتم تفسير 81٪ من المتغيرات التابعة (قيم y) بواسطة المتغيرات المستقلة (قيم x). بشكل عام ،

Observations:

174=وهو عدد المشاهدات المدخلة في النموذج

الجزء الثاني من الناتج هو تحليل التباين (ANOVA): نادراً ما يتم استخدام جزء ANOVA لتحليل انحدار خطي بسيط في Excel

لكن لا باس في الإطلاع على النتائج:

df هو عدد درجات الحرية المرتبطة بمصادر التباين.

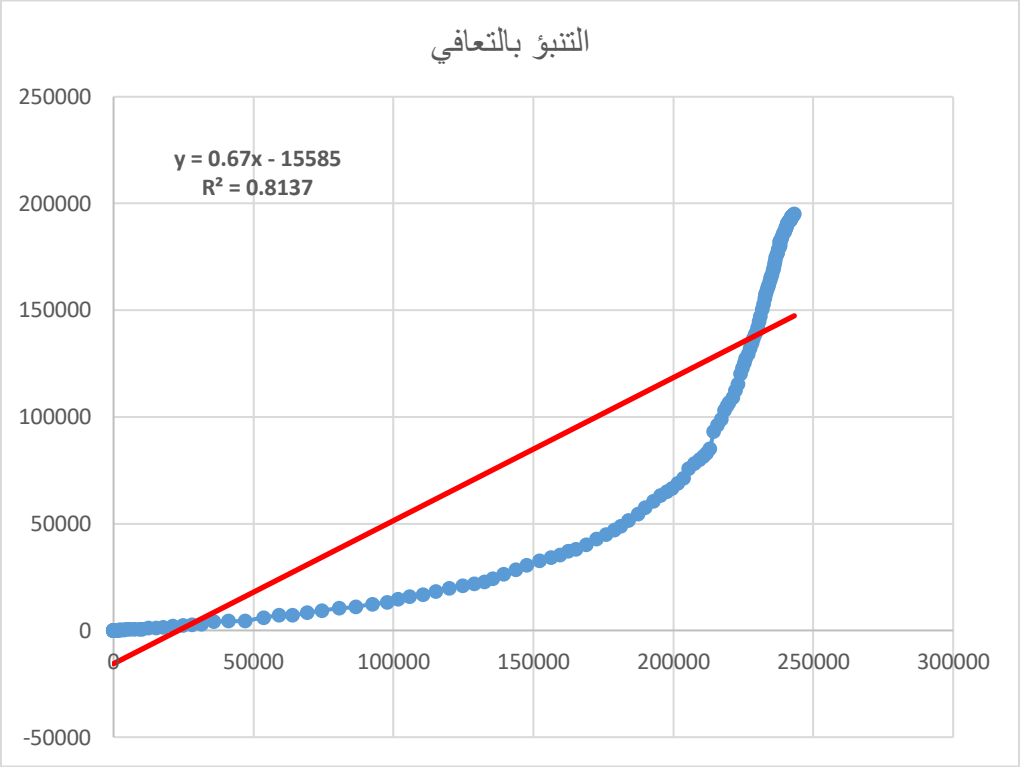
SS هو مجموع المربعات. كلما صغر حجم SS المتبقي مقارنة بـ Total SS ، كان نموذجنا يناسب البيانات بشكل أفضل.

وعليه نلاحظ انا SS هو اصغر من ss Total أي ان النموذج يتناسب مع البيانات

MS هو مربع الوسط

Significance F الخاصة بالنموذج هي أقل من 0.05(5%) فإننا نستنتج اننا لسنا في حاجة لتغيير المتغير المستقل

المخطط البياني :



يتم صياغة معادلة الانحدار الخطي رياضيا : $y = bx + a + \varepsilon$

حيث أن

- x متغير مستقل (التعافي)
- y متغير تابع (الإصابات)

ملاحظة

a هي تقاطع Y ، وهي القيمة المتوسطة المتوقعة لـ y عندما تكون جميع المتغيرات x تساوي 0. في الرسم البياني للانحدار ، هي النقطة التي يتقاطع فيها الخط مع المحور Y.

b هو ميل خط الانحدار ، وهو معدل التغيير لـ y مع تغير x.

ε هو مصطلح الخطأ العشوائي ، وهو الفرق بين القيمة الفعلية لمتغير تابع وقيمته المتوقعة.

يتم حساب الخطأ المعيار تلقائيا في برنامج اكسل

لتكون المعادلة بالشكل النهائي : $y = bx + a$

$$Injuries = b * recovered + a$$

بالتعويض في معادية المستقيم : $y = 0.67x - 15585$

	7/7/2020	241956	192815
	7/8/2020	242149	193640
	7/9/2020	242363	193978
	7/10/2020	242639	194273
	7/11/2020	242827	194579
	7/12/2020	243061	194928
	7/13/2020	243230	195106
	7/14/2020	243344	147455.5
y = 0.67x - 15585			
R² = 0.8137			

وعليه نعوض المعادلة المستخرجة من تحليل الانحدار بقيمة البيانات الخاصة بعدد الإصابات بتاريخ 7/14/2020

حتى نتنبأ بعدد المصابين المتوقع تعافيهم في هذا اليوم

$$y = 0.67 * 243344 - 15585$$

147455 = وعليه يمكننا معرفة عدد الأشخاص الذي من المتوقع تعافيهم عند الإصابة بالفيروس في تاريخ 7/14/2020

أما بالنسبة للعدد الفعلي الذي حدث في ذاك اليوم فكان =195441 وهو رقم قد يكون التفسير الأصح له هو السلوك الغير قابل للتنبؤ بدقة للفيروس, كما انها فتره قد بدأت فيه ايطاليا في وصولها إلى خط الذروة كما سنشاهد في القسم العملي لنموذج النمو اللوجستي

والسبب في وجود فارق بين القيمة المتوقعة والقيمة الحقيقية هو ان تحليل الإنحدار يعطي قيمة متوقعة وليست دقيقة للمعيار الناتج عن مجموعة خطية من المتنبئات.

كما بين النموذج فإن العلاقة إيجابية: إذ أظهر الخط اتجاها ً تصاعديا ً. أي مع زيادة المتغير المستقل ، يزداد المتغير التابع أيض ًا.

Syria

I – الإصابات والأموات

المتغير التابع المراد التنبؤ به : الأموات (عدد الأموات)

المتغيرات المستقلة الذي يؤثر على المتغير التابع : الإصابات (عدد الإصابات)

بعد إجراء التحليل خلصت النتائج إلى التالي :

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.974467377
R Square	0.949586668
Adjusted R Square	0.949293568
Standard Error	0.873290558
Observations	174

ANOVA

	df	SS	MS	F	Significance F
Regression	1	2470.78631	2470.78631	3239.795941	1.6351E-113
Residual	172	131.1734606	0.762636399		
Total	173	2601.95977			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	0.512154774	0.081331442	6.297131331	2.44669E-09	0.35161853	0.672691017	0.35161853	0.672691017
X Variable 1	0.035131782	0.000617222	56.91920538	1.6351E-113	0.033913477	0.036350087	0.033913477	0.036350087

التفسير:

Multiple R:

إن قوة العلاقة الخطية بين متغيرين هي 0.974467377 اي انها قريبة من الواحد (1)

ومنه نستنتج أن العلاقة الخطية بين المتغيرين علاقة قوية وقد تكون قوية جدا

R Square :

معامل التحديد في مثالنا السابق 0.949586668

في مثالنا ، R2 هو **0.94** ، وهو أمر جيد. هذا يعني أن 94٪ من قيمنا تناسب نموذج تحليل الانحدار. بمعنى آخر ، يتم تفسير 94٪ من المتغيرات التابعة (قيم y) بواسطة المتغيرات المستقلة (قيم x). بشكل عام .

Observations :

174=وهو عدد المشاهدات المدخلة في النموذج

الجزء الثاني من الناتج هو تحليل التباين (ANOVA): نادرًا ما يتم استخدام جزء ANOVA لتحليل انحدار خطي بسيط في Excel

ولكن لا باس في الإطلاع على النتائج

df هو عدد درجات الحرية المرتبطة بمصادر التباين.

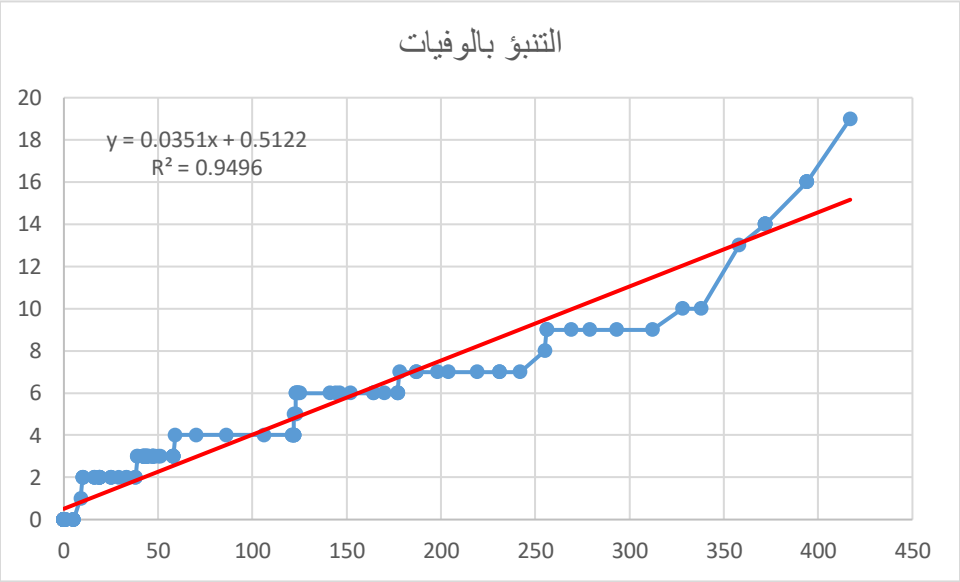
SS هو مجموع المربعات. كلما صغر حجم SS المتبقي مقارنة بـ Total SS ، كان نموذجنا يناسب البيانات بشكل أفضل.

وعليه نلاحظ انا SS هو اصغر من ss Total أي ان النموذج يتناسب مع البيانات

MS هو مربع الوسط

Significance F الخاصة بالنموذج هي أقل من 0.05(5%) وعيله فإننا نستنتج اننا لسنا في حاجة لتغيير المتغير المستق في حاكل كانت Significance F اكبر من 0.05فيجب تغيير المتغير المسل

المخطط البياني :



يتم صياغة معادلة الانحدار الخطي رياضيا $y = bx + a + \varepsilon$

حيث أن

- **x** متغير مستقل (الموتى)
- **y** متغير تابع (الإصابات)

ملاحظة

a هي تقاطع Y ، وهي القيمة المتوسطة المتوقعة لـ y عندما تكون جميع المتغيرات x تساوي **0**. في الرسم البياني للانحدار ، هي النقطة التي يتقاطع فيها الخط مع المحور Y.

b هو ميل خط الانحدار ، وهو معدل التغيير لـ y مع تغير x.

ε هو مصطلح الخطأ العشوائي ، وهو الفرق بين القيمة الفعلية لمتغير تابع وقيمته المتوقعة.

يتم حساب الخطأ المعيار تلقائيا في برنامج اكسل

لتكون المعادلة بالشكل النهائي : $y = bx + a$

$$Injuries = b * The\ dead + a$$

بالتعويض في معادية المستقيم : $y = 0.0351x + 0.5122$

169	7/7/2020	372	14
170	7/8/2020	372	14
171	7/9/2020	372	14
172	7/10/2020	394	16
173	7/11/2020	394	16
174	7/12/2020	394	16
175	7/13/2020	417	19
176	7/14/2020	439	15.9211
177			
178			

وعليه نعوض المعادلة المستخرجة من تحليل الانحدار بقيمة البيانات الخاصة بعدد الإصابات بتاريخ 7/14/2020

حتى نتنبأ بعدد الأموات المتوقع وقوعها في هذا اليوم

$y = 0.0351 * 439 + 0.5122$

15.9 = تقريبيا = 16 وعليه يمكننا معرفة عدد الأشخاص الذي من المتوقع موتهم عند الإصابة بالفيروس في تاريخ 7/14/2020

أما بالنسبة للعدد الفعلي الذي حدث في ذاك اليوم فكان = 20 وهي قيمة قريبة للنسبة المتوقع موتها

والسبب في وجود فارق بين القيمة المتوقعة والقيمة الحقيقية هو ان تحليل الإنحدار يعطي قيمة متوقعة وليست دقيقة للمعيار الناتج عن مجموعة خطية من المتنبئات.

كما هو مبين في الرسم البياني فإن العلاقة إيجابية: إذ أظهر الخط اتجاها تصاعديا .
هذا يشير إلى أنه مع زيادة المتغير المستقل ، يزداد المتغير التابع أيضا.

2– الإصابات والتعافي

المتغير التابع المراد التنبؤ به : المتعافي (عدد المتعافيين)

المتغيرات المستقلة الذي يؤثر على المتغير التابع : الإصابات (عدد الإصابات)

بعد إجراء التحليل خلصت النتائج إلى التالي :

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.9836407
R Square	0.967549027
Adjusted R Square	0.967360359
Standard Error	7.140549872
Observations	174

ANOVA

	df	SS	MS	F	Significance F
Regression	1	261479.1237	261479.1237	5128.303357	5.7044E-130
Residual	172	8769.841826	50.98745248		
Total	173	270248.9655			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	1.958226039	0.665014885	2.9446349	0.00368079	0.645584966	3.270867113	0.645584966	3.270867113
X Variable 1	0.361410773	0.005046778	71.61217325	5.7044E-130	0.351449178	0.371372367	0.351449178	0.371372367

التفسير:

Multiple R:

إن قوة العلاقة الخطية بين متغيرين هي 0.98364078 اي انها قريبة من الواحد (1)

ومنه نستنتج أن العلاقة الخطية بين المتغيرين علاقة قوية نوعا ما

R Square :

معامل التحديد في مثالنا السابق 0.967549027

في مثالنا ، R2 هو 0.96، وهو أمر جيد نوعاً ما. هذا يعني أن 96٪ من قيمنا تناسب نموذج تحليل الانحدار. بمعنى آخر ، يتم تفسير 96٪ من المتغيرات التابعة (قيم y) بواسطة المتغيرات المستقلة (قيم x). بشكل عام ،

Observations :

174= وهو عدد المشاهدات المدخلة في النموذج

الجزء الثاني من الناتج هو تحليل التباين (ANOVA): نادراً ما يتم استخدام جزء ANOVA لتحليل انحدار خطي بسيط في Excel

ولكن لا باس في الإطلاع على النتائج:

df هو عدد درجات الحرية المرتبطة بمصادر التباين.

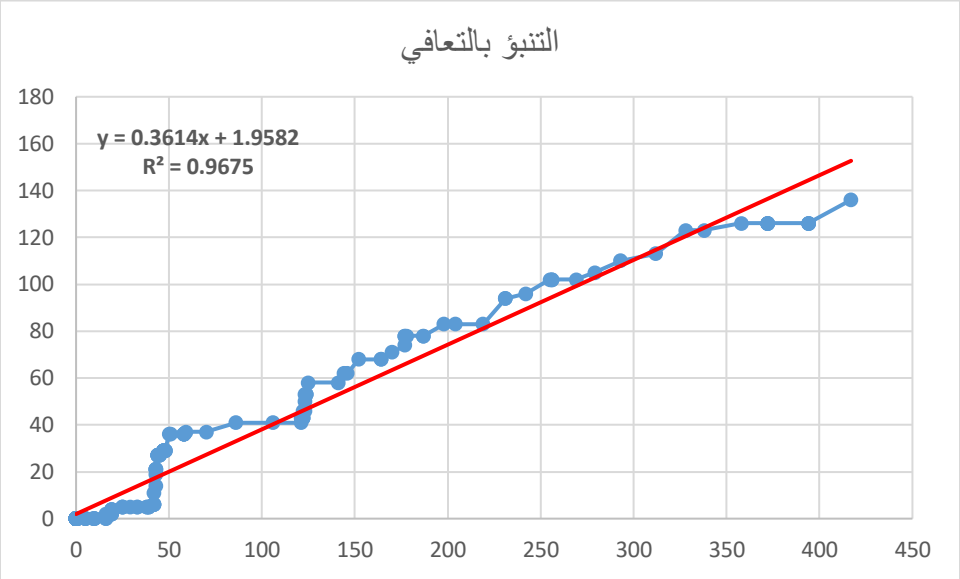
SS هو مجموع المربعات. كلما صغر حجم SS المتبقي مقارنة بـ Total SS ، كان نموذجنا يناسب البيانات بشكل أفضل.

وعليه نلاحظ انا ss هو اصغر من Total ss أي ان النموذج يتناسب مع البيانات

MS هو مربع الوسط

Significance F الخاصة بالنموذج هي أقل من 0.05(5%) وعيله فإننا نستنتج اننا لسنا في حاجة لتغيير المتغير المستق في حاكل كانت Significance F اكبر من 0.05فيجب تغيير المتغير المسل

المخطط البياني :



يتم صياغة معادلة الانحدار الخطي رياضيا: $y = bx + a + \varepsilon$

حيث أن

- x متغير مستقل (التعافي)
- y متغير تابع (الإصابات)

ملاحظة

a هي تقاطع Y ، وهي القيمة المتوسطة المتوقعة لـ y عندما تكون جميع المتغيرات x تساوي 0. في الرسم البياني للانحدار ، هي النقطة التي يتقاطع فيها الخط مع المحور Y.

b هو ميل خط الانحدار ، وهو معدل التغيير لـ y مع تغير x.

ε هو مصطلح الخطأ العشوائي ، وهو الفرق بين القيمة الفعلية لمتغير تابع وقيمته المتوقعة.

يتم حساب الخطأ المعيار تلقائيا خلف في برنامج اكسل

لتكون المعادلة بالشكل النهائي :

$$y = bx + a$$

$$Injuries = b * recovered + a$$

بالتعويض في معادية المستقيم : $y = 0.3614x + 1.9582$

7/7/2020	372	126
7/8/2020	372	126
7/9/2020	372	126
7/10/2020	394	126
7/11/2020	394	126
7/12/2020	394	126
7/13/2020	417	136
7/14/2020	439	160.6128

وعليه نعوض المعادلة المستخرجة من تحليل الانحدار بقيمة البيانات الخاصة بعدد الإصابات بتاريخ 7/14/2020

حتى نتنبأ بعدد الإصابات المتوقع تعافيها في هذا اليوم

$y = 0.3614 * 439 + 1.9582$

= 160 وعليه يمكننا معرفة عدد الأشخاص الذي من المتوقع تعافيهم عند الإصابة بالفيروس في تاريخ 7/14/2020

أما بالنسبة للعدد الفعلي الذي حدث في ذاك اليوم فكان = 138 وهي قيمة قريبة بشكل قليل للقيمة الحقيقة وذلك قد يكون راجعا إلى التشتت في طريقة إنتشار الفيروس

والسبب في وجود فارق بين القيمة المتوقعة والقيمة الحقيقية هو ان تحليل الإنحدار يعطي قيمة متوقعة وليست دقيقة للمعيار الناتج عن مجموعة خطية من المتنبئات.

كما هو مبين في الرسم البياني فإن العلاقة إيجابية: إذ أظهر الخط اتجاها تصاعديا . هذا يشير إلى أنه مع زيادة المتغير المستقل ، يزداد المتغير التابع أيضا

South Korea

1 – الإصابات والأموات

المتغير التابع المراد التنبؤ به : الأموات (عدد الأموات)

المتغيرات المستقلة الذي يؤثر على المتغير التابع : الإصابات (عدد الإصابات)

بعد إجراء التحليل خلصت النتائج إلى التالي :

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.941728413
R Square	0.886852404
Adjusted R Square	0.886194569
Standard Error	38.50756095
Observations	174

ANOVA

	df	SS	MS	F	Significance F
Regression	1	1999063.02	1999063.02	1348.138347	2.64728E-83
Residual	172	255047.1471	1482.832251		
Total	173	2254110.167			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	-24.09060073	6.010840463	-4.007858947	9.11217E-05	-35.95511122	-12.22609024	-35.95511122	-12.22609024
X Variable 1	0.02309921	0.000629115	36.71700352	2.64728E-83	0.02185743	0.024340989	0.02185743	0.024340989

التفسير:

Multiple R:

إن قوة العلاقة الخطية بين متغيرين هي 0.941728413 اي انها قريبة من الواحد (1)

ومنه نستنتج أن العلاقة الخطية بين المتغيرين علاقة قوية

R Square:

معامل التحديد في مثالنا السابق 0.886852404

في مثالنا ، R2 هو 0.88، وهو أمر جيد. هذا يعني أن 88٪ من قيمنا تناسب نموذج تحليل الانحدار. بمعنى آخر ، يتم تفسير 88٪ من المتغيرات التابعة (قيم y) بواسطة المتغيرات المستقلة (قيم x). بشكل عام ،

Observations

174 = وهو عدد المشاهدات المدخلة في النموذج

الجزء الثاني من الناتج هو تحليل التباين (ANOVA): نادرًا ما يتم استخدام جزء ANOVA لتحليل انحدار خطي بسيط في Excel

ولكن لا باس في الإطلاع على النتائج

df هو عدد درجات الحرية المرتبطة بمصادر التباين.

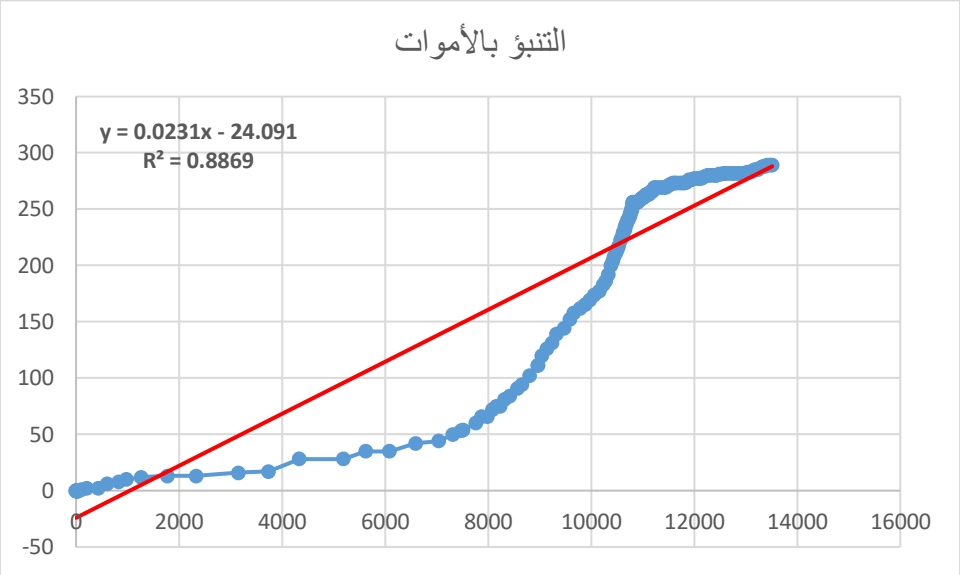
SS هو مجموع المربعات. كلما صغر حجم SS المتبقي مقارنة بـ Total SS ، كان نموذجنا يناسب البيانات بشكل أفضل.

وعليه نلاحظ انا ss هو اصغر من Total ss أي ان النموذج يتناسب مع البيانات

MS هو مربع الوسط

Significance F الخاصة بالنموذج هي أقل من 0.05(5%) وعيله فإننا نستنتج اننا لسنا في حاجة لتغيير المتغير المستق في حاكل كانت Significance F اكبر من 0.05فيجب تغيير المتغير المسل

المخطط البياني :



يتم صياغة معادلة الانحدار الخطي رياضيا: $y = bx + a + \varepsilon$

حيث أن

- **x** متغير مستقل (الموتى)
- **y** متغير تابع (الإصابات)

ملاحظة

a هي تقاطع Y ، وهي القيمة المتوسطة المتوقعة لـ y عندما تكون جميع المتغيرات x تساوي 0. في الرسم البياني للانحدار ، هي النقطة التي يتقاطع فيها الخط مع المحور Y.

b هو ميل خط الانحدار ، وهو معدل التغيير لـ y مع تغير x.

ε هو مصطلح الخطأ العشوائي ، وهو الفرق بين القيمة الفعلية لمتغير تابع وقيمته المتوقعة.

يتم حساب الخطأ المعيار تلقائيا خلف في برنامج اكسل

لتكون المعادلة بالشكل النهائي :

$$y = bx + a$$

$$Injuries = b * The\ dead + a$$

بالتعويض في معادية المستقيم : $y = 0.0231x - 24.091$

وعليه نعوض المعادلة المستخرجة من تحليل الانحدار بقيمة البيانات الخاصة بعدد الإصابات بتاريخ 7/14/2020

حتى نتنبأ بعدد الأموات المتوقع وقوعها في هذا اليوم

$y = 0.0231 * 13551 - 24.091$

169	7/7/2020	13244	285
170	7/8/2020	13293	287
171	7/9/2020	13338	288
172	7/10/2020	13373	288
173	7/11/2020	13417	289
174	7/12/2020	13479	289
175	7/13/2020	13512	289
176	7/14/2020	13551	288.9371
177			
178			

288.9 = تقريبيا = 289 وعليه يمكننا معرفة عدد الأشخاص الذي من المتوقع موتهم عند الإصابة بالفيروس في تاريخ 7/14/2020

أما بالنسبة للعدد الفعلي الذي حدث في ذاك اليوم فكان = 288 وهي قيمة مطابق للنسبة المتوقع موتها

وعليه فإن النموذج اعطى قيمة تنبؤ دقيقة رغم أن تحليل الإنحدار يعطي قيمة متوقعة وليست دقيقة للمعيار الناتج عن مجموعة خطية من المتنبئات.

كما هو مبين في الرسم البياني فإن العلاقة إيجابية: إذ أظهر الخط اتجاها تصاعديا .

هذا يشير إلى أنه مع زيادة المتغير المستقل ، يزداد المتغير التابع أيضا.

2– الإصابات والتعافي

المتغير التابع المراد التنبؤ به : المتعافي (عدد المتعافيين)

المتغيرات المستقلة الذي يؤثر على المتغير التابع : الإصابات (عدد الإصابات)

بعد إجراء التحليل خلصت النتائج إلى التالي :

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.895854328
R Square	0.802554977
Adjusted R Square	0.801407041
Standard Error	2091.820653
Observations	174

ANOVA

	df	SS	MS	F	Significance F
Regression	1	3059186409	3059186409	699.1286	1.73407E-62
Residual	172	752622747.1	4375713.646		
Total	173	3811809156			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	-1401.267226	326.5228935	-4.291482323	2.96E-05	-2045.775147	-756.7593037	-2045.775147	-756.7593037
X Variable 1	0.903622104	0.034174984	26.44103944	1.73E-62	0.836165737	0.971078471	0.836165737	0.971078471

التفسير:

Multiple R:

إن قوة العلاقة الخطية بين متغيرين هي **0.895854328** اي انها قريبة قليلا من الواحد (1) ومنه نستنتج أن العلاقة الخطية بين المتغيرين علاقة جيدة

R Square :

معامل التحديد في مثالنا السابق 0.802554977

في مثالنا ، R2 هو 0.80، وهو أمر جيد نوعا ما. هذا يعني أن 80٪ من قيمنا تناسب نموذج تحليل الانحدار. بمعنى آخر ، يتم تفسير 80٪ من المتغيرات التابعة (قيم y) بواسطة المتغيرات المستقلة (قيم **x**). بشكل عام ،

Observations:

174 = وهو عدد المشاهدات المدخلة في النموذج

الجزء الثاني من الناتج هو تحليل التباين (ANOVA): نادراً ما يتم استخدام جزء ANOVA لتحليل انحدار خطي بسيط في Excel

ولكن لا باس في الإطلاع على النتائج

df هو عدد درجات الحرية المرتبطة بمصادر التباين.

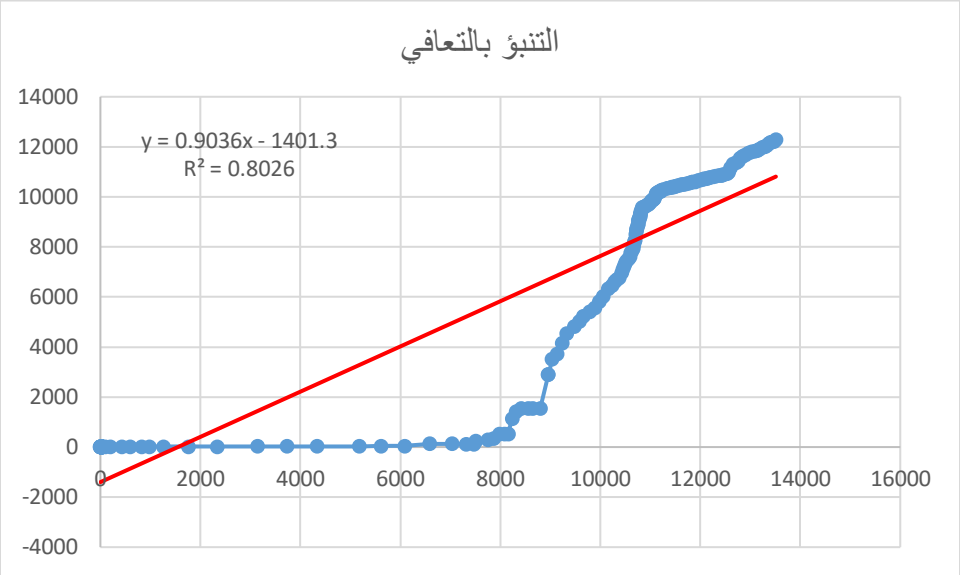
SS هو مجموع المربعات. كلما صغر حجم SS المتبقي مقارنة بـ Total SS ، كان نموذجنا يناسب البيانات بشكل أفضل.

وعليه نلاحظ انا SS هو اصغر من ss Total أي ان النموذج يتناسب مع البيانات

MS هو مربع الوسط

Significance F الخاصة بالنموذج هي أقل من 0.05(5%) وعيله فإننا نستنتج اننا لسنا في حاجة لتغيير المتغير المستق في حاكل كانت Significance F اكبر من 0.05فيجب تغيير المتغير المسل

المخطط البياني :



يتم صياغة معادلة الانحدار الخطي رياضيا $y = bx + a + \varepsilon$

حيث أن

- x متغير مستقل (التعافي)
- y متغير تابع (الإصابات)

ملاحظة

a هي تقاطع Y ، وهي القيمة المتوسطة المتوقعة لـ y عندما تكون جميع المتغيرات x تساوي 0. في الرسم البياني للانحدار ، هي النقطة التي يتقاطع فيها الخط مع المحور Y.

b هو ميل خط الانحدار ، وهو معدل التغيير لـ y مع تغير x.

ε هو مصطلح الخطأ العشوائي ، وهو الفرق بين القيمة الفعلية لمتغير تابع وقيمته المتوقعة.

يتم حساب الخطأ المعيار تلقائيا خلف في برنامج اكسل

لتكون المعادلة بالشكل النهائي :

$$y = bx + a$$

$$Injuries = b * recovered + a$$

بالتعويض في معادية المستقيم : $y = 0.9036x - 1401.3$

7/7/2020	13244	11970
7/8/2020	13293	12019
7/9/2020	13338	12065
7/10/2020	13373	12144
7/11/2020	13417	12178
7/12/2020	13479	12204
7/13/2020	13512	12282
7/14/2020	13551	10843.38

وعليه نعوض المعادلة المستخرجة من تحليل الانحدار بقيمة البيانات الخاصة بعدد الإصابات بتاريخ 7/14/2020 حتى نتنبأ بعدد الإصابات المتوقع تعافيها في هذا اليوم

$y = 0.9036 * 13551 - 1401.3$

$= 10843.3$ وعليه يمكننا معرفة عدد الأشخاص الذي من المتوقع تعافيهم عند الإصابة بالفيروس في تاريخ 7/14/2020

أما بالنسبة للعدد الفعلي الذي حدث في ذاك اليوم فكان =12348

وهي قيمة أقل من القيمة الحقيقية بنسبة قليلة وذلك قد يكون راجعاً إلى التشتت في طريقة إنتشار الفيروس

والسبب في وجود فارق بين القيمة المتوقعة والقيمة الحقيقية هو ان تحليل الإنحدار يعطي قيمة متوقعة وليست دقيقة للمعيار الناتج عن مجموعة خطية من المتنبئات.

كما هو مبين في الرسم البياني فإن العلاقة إيجابية: إذ أظهر الخط اتجاهاً تصاعدياً .
هذا يشير إلى أنه مع زيادة المتغير المستقل ، يزداد المتغير التابع أيضاً.

المقاربة الثانية: إقتراح نموذج للتنبؤ بإتجاه إنتشار المرض والتنبؤ بعدد الإصابات التي ستع بفترات مستقبلية.

التطبيق العملي باستخدام : Time seies

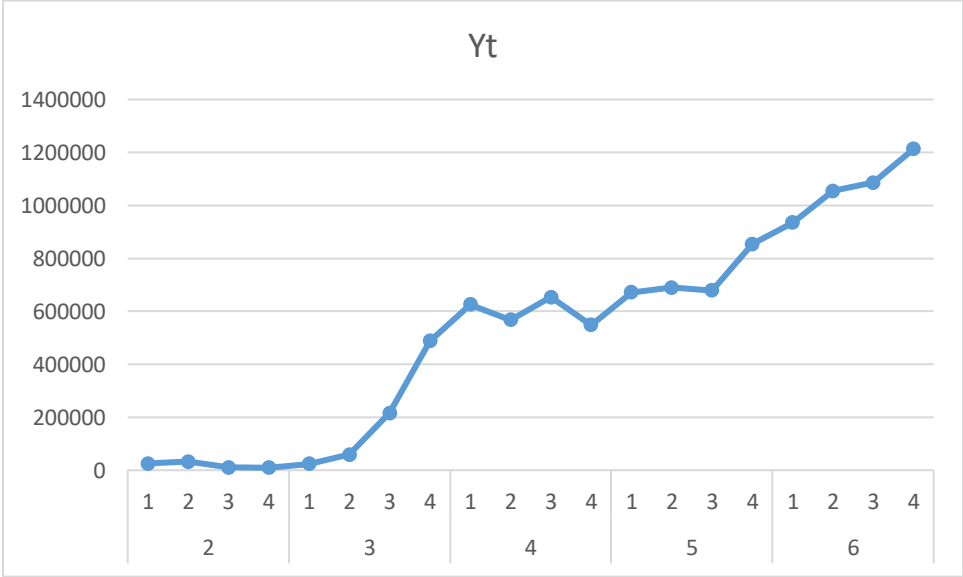
سيتم تحليل البيانات السابقة الخاصة بإصابات فيروس كورونا خلال تسلسل مرتب للبيانات على مسافات زمنية متساوية سابقة قدرها 5 أشهر تمتد من الشهر الثاني إلى الشهر السادس وذلك للتنبؤ بمستقبل عدد الإصابات خلال الشهر السابع 7 وذلك لعرض الإتجاه العام لإنتشار الفيروس + القدرة على أخذ الحذر خلال الفترات السابقة عن طريق التوقع لعدد الإصابات خلال فترة زمنية مستقبلية وتهيئة المراكز والمشافي بالاستعداد التام .

اعتمد الباحث طريقة المتوسط المتحرك moving average في نمذجة السلسلة الزمنية .

اما بالنسبة للبلدان التي تم اجراء المسح لها فلقد جمع الباحث عدد البلدان وهم موزعين عالشكل التالي مقدرين ب191 بلد كما في الجدول التالي .

Botswana	Zimbabwe	Suriname	Portugal	Mongolia	Kazakhstan	Germany	Cuba	Australia	Afghanistan
Burundi	Canada	Sweden	Qatar	Montenegro	Kenya	Ghana	Cyprus	Austria	Albania
Sierra Leone	Dominica	Switzerland	Romania	Morocco	Korea, South	Greece	Czechia	Azerbaijan	Algeria
Netherlands	Grenada	Taiwan*	Russia	Namibia	Kuwait	Guatemala	Denmark	Bahamas	Andorra
Malawi	Mozambique	Tanzania	Rwanda	Nepal	Kyrgyzstan	Guinea	Djibouti	Bahrain	Angola
United Kingdom	Syria	Thailand	Saint Lucia	Netherlands	Latvia	Haiti	Dominican Republic	Bangladesh	Antigua and Barbuda
France	Timor-Leste	Togo	Saint Vincent and the Grenadines	New Zealand	Lebanon	Holy See	Ecuador	Barbados	Argentina
South Sudan	Belize	Trinidad and Tobago	San Marino	Nicaragua	Liberia	Honduras	Egypt	Belarus	Armenia
Western Sahara	Laos	Tunisia	Saudi Arabia	Niger	Liechtenstein	Hungary	El Salvador	Canada	Belgium
Sao Tome and Principe	Libya	Turkey	Senegal	Nigeria	Lithuania	Iceland	Equatorial Guinea	Central African Republic	Benin
Yemen	West Bank and Gaza	Uganda	Serbia	North Macedonia	Luxembourg	India	Eritrea	Chad	Bhutan
	Guinea-Bissau	Ukraine	Seychelles	Norway	Madagascar	Indonesia	Estonia	Chile	Bolivia
	Mali	United Arab Emirates	Singapore	Oman	Malaysia	Iran	Eswatini	China	Bosnia and Herzegovina
	Saint Kitts and Nevis	United Kingdom	Slovakia	Pakistan	Maldives	Iraq	Ethiopia	Colombia	Brazil
	Canada	Uruguay	Slovenia	Panama	Malta	Ireland	Fiji	Congo (Brazzaville)	Brunei
	Canada	US	Somalia	Papua New Guinea	Mauritania	Israel	Finland	Congo (Kinshasa)	Bulgaria
	Kosovo	Uzbekistan	South Africa	Paraguay	Mauritius	Italy	France	Costa Rica	Burkina Faso
	Comoros	Venezuela	Spain	Peru	Mexico	Jamaica	Gabon	Cote d'Ivoire	Cabo Verde
	Tajikistan	Vietnam	Sri Lanka	Philippines	Moldova	Japan	Gambia	Croatia	Cambodia
	Lesotho	Botswana	Sudan	Poland	Monaco	Jordan	Georgia	Diamond Princess	Cameroon

بداية تم إجراء التحليل الاستكشافي والذي من خلاله تم عرض نخطط لعدد الإصابات مع مرور الوقت خلال 5 أشهر مقسمة مقسم كل شهر إلى ربع أسابيع.



ملاحظة للمخطط القيم في الشهر الأول ليست 0 ولكن لصعوبة عرض المخطط لكافة القيم المتغيرة المتعلقة بالإصابات لكثرة عددها يكتفي بعرض التباين العالي بين القيم

TREND: الإتجاه لقد لوحظ نمط الزيادة على امتداد الفترات الزمنية كافة . وعلى هذه الحالة كان الاتجاه الأساسي المتزايد تدريجيًا. أي ازداد عدد الإصابات على مدى فترة من الزمن.

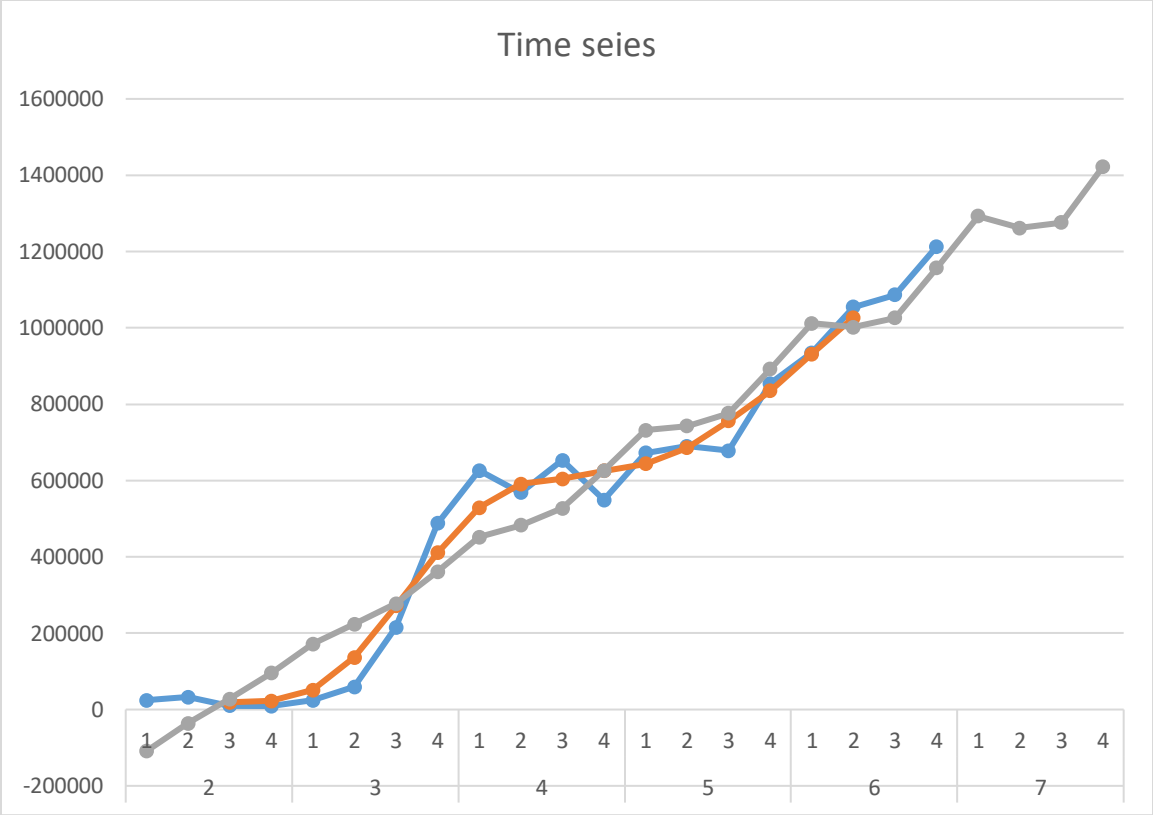
حيث كما تم ملاحظته ان الإنتشار في الشهر الثاني كان خفيف نوعًا ما مقارنة بباقي الأشهر لننطلق إلى إرتفاع اسي مباشر إلى الشهر الثالث وهو يعني زيادة في عدد

الإصابات في الشهر الثالث عالمياً , ليحافظ على معدل تقريباً يمكن وصفه بالثابت ويمتد الشهر الرابع في الأسبوع الثاني إلى منتصف الشهر الخامس , ليعاود الزيادة التدريجية في عدد الإصابات إلى نهاية الشهر السادس

- خلا الأشهر ال5 التي تم إجراء الدراسة عليها كانت النتائج المترتبة لمجموع الإصابات موزعة وفق الأتي:

Injuries	Quarter	Month
24468	1	2
32513	2	
9932	3	
9169	4	
23951	1	3
58981	2	
215479	3	
487556	4	
626190	1	4
568135	2	
652897	3	
548738	4	
671962	1	5
689740	2	
678462	3	
854309	4	
934876	1	6
1054374	2	
1086166	3	
1212631	4	
???	1	7
???	2	
???	3	
???	4	

العدد المسند للإصابات خلال الاسابيع وضمن الأشهر في البيانات السابقة هو العدد الفعلي وليس التجميع التراكمي



يعرض المخطط التالي الخاص بنموذج لسلاسل الزمنية الإصابات التي حصلت خلال عام 2020 موزعة عبر 13 اشهر (2.3.4.5.6) وتعرض التنبؤ بعدد الإصابات الخاص بالشهر السابع من 2020 .. كما يعرض المخطط الإتجاه العام لإنتشار المرض .

كما يبين النموذج .. خلال الأشهر ال**2** و **3** و **4** نلاحظ الفروقات الواضحة بين القيم الحقيقية والقيم المتنبئ بها لنصل إلى الشهر ال**5** و **6** لتصبح القيم متقاربة جزئيا .. ومن هذا نستنتج التالي:

بأن وجود تذبذب وعدم تطابق واضح بين القيم الحقيقية والقيم التي تم التنبؤ بها يعطي بأن الفيروس لا يمكن التنبؤ بآلية إنتشاره وان الإنتشار لا يسير وفق آلية منتظمة .

- بيانات الشهر السابع المتوقع حدوثها على امتداد 4 اسابيع

الشهر السابع		
	1	1292399
	2	1261707
	3	1275619
	4	1422154

نلاحظ خلال الأربعة أسابيع المتنبأ بها في الشهر السابع لارتفاع عدد الإصابات في الأسبوع الأول مقارنة مع ما سبق من الاشهر **2** و **3** و **4** و **5** و **6** ثم نلاحظ إنخفاض في عدد الإصابات في الأسبوع الثاني والثالث لنشهد بعدها ارتفاع حاد في عدد الإصابات في الأسبوع الرابع بفارق كبير.

يمكن الإستفادة من النتائج التالية عبر التي:

يمكننا معرفة كيف يتغير وينتشر الفيروس بمرور الوقت وعليه بما يصحبه من عدد الإصابات.

يمكننا استخدام نموذج السلسلة الزمنية أيضًا لفحص كيفية مقارنة التغييرات المرتبطة بنقطة البيانات المختارة بالتغيرات في المتغيرات الأخرى خلال نفس الفترة الزمنية.

كما يمكننا استخدام النتائج للأشهر **2.3.4.5.6**.وايضا **7** للتنبؤ بالفترات الزمنية القادمة لعدد الإصابات المتوقع حدوثة

المقاربة الثالثة: إقتراح نموذج لملاحظة نسبة نمو الفيروس في مجموعة من البلدان وذلك لوضع مقياس للبلدان التي قد وصل عدد الإصابات فيها إلى الذروة

التطبيق العملي باستخدام : logistic growth model

باستخدام نموذج النمو اللوجستي

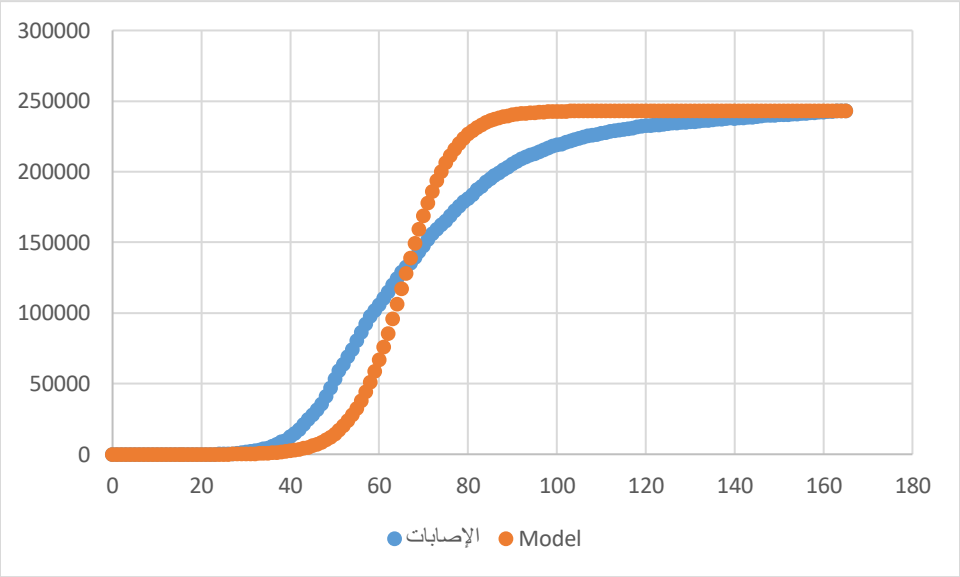
والسبب في استخدام الباحث لنموذج النمو اللوجستي لنمذجة تفشي فيروس كورونا , هو أن علماء الأوبئة قد درسوا تلك الأنواع من الفاشيات ومن المعروف جيدًا أن الفترة الأولى للوباء تتبع النمو الأسي كما شهدنا في تحليل الإنحدار وأن الفترة الإجمالية يمكن نمذجتها بنمو لوجستي.

تم تحليل البيانات الخاصة بإصابات فيروس كورونا لكل من وايطاليا , سوريا , كوريا الجنوبية وذلك خلال تسلسل مرتب للبيانات يبدأ من تاريخ 2/1/2020 إلى تاريخ 7/14/2020

وذلك باستخدام نموذج النوم اللوجستي حيث قدم لنا النموذج عرض توضيحي رياضي بالبلدان التي قد وصلت إلى حد الذروة في عدد الإصابات وقد عرض لنا الزيادة في النمو للفترة البداية ، ولكن النمو المتناقص في مرحلة لاحقة وهو كلما اقتربنا من الحد الأقصى

وهو عكس النموذج الأسي الذي يبقى متصاعدا .. يقدم لنا النوذج اللوجستي دراسة للفترة الزمنية الحالية ويقدر رياضيًا هل قد وصلت عدد الإصابات إلى حد الذروة وفق البيانات التاريخية أم ان النموذج لهذا البلد مازال مستمر ولو يصل للحد

Italia / ايطاليا



تم إجراء النموذج على فترة زمنية تبدأ من تاريخ 2/1/2020 إلى تاريخ 7/14/2020 وقدرها 165 يوم

ملاحظة للنموذج :

القيم في ال٤٠ يوم الاولى ليست 0 ولكن لصعوبة عرض المخطط لكافة القيم المتغيرة المتعلقة بالإصابات لكثرة عددها يكتفي بعرض التباين العالي بين القيم .

فكما نلاحظ في النموذج فقد بين لنا بداية ارتفاع في عدد الإصابات كنمو اسي ليستقر بعدا بوضع شبه ساكن في الزيادة وهو ما ندعوه في نموذج النمو اللوجستي بحد الذروة وبمثالنا هنا نستطيع أن نتوقع بأن أيطاليا قد وصلت إلى حد الذروة كما هو مبين في النموذج لتكون حد الذروة للإصابات بمدينة ايطاليا = 250000 إصابة ..

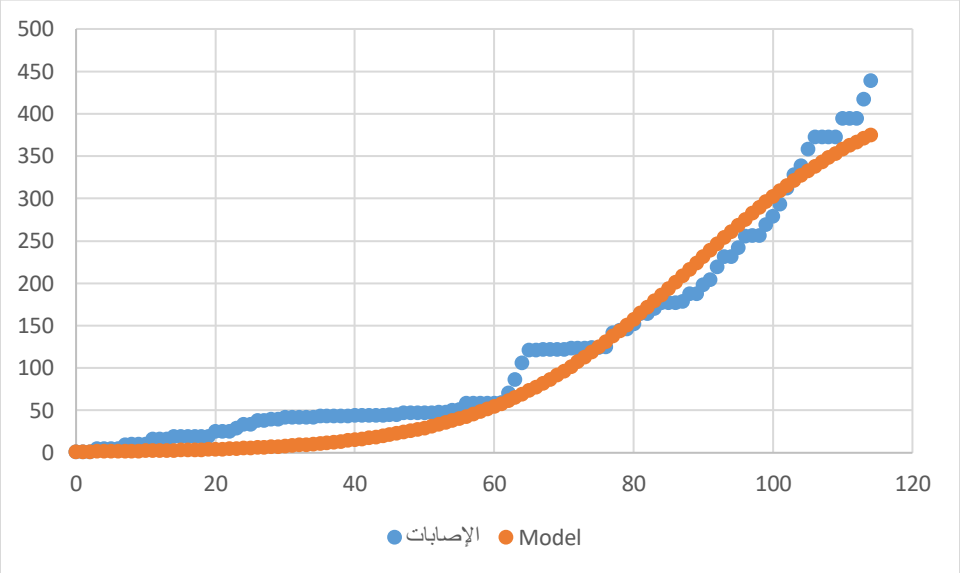
وهو يتناسب طرداً بما تم التصريح به من قبل الحكومة الإيطالية بوصولها إلى حد الذروة في الإصابات

ومنه نستنتج : أستطاعت ايطاليا السيطرة أو الحد من إنتشار الفيروس على الرغم من وجود فترات أولية قد عانت منها من التفشي الأسي

Syri a /سوريا

نستعرض النموذج اللوجستي الخاص بسوريا , وبسبب لأن النمو مايزال في مراحله الاولى ضمن المؤكد ان ما سيستعرضه المنحني اللوجستي هو مختلف بدرجات كبيرة عما تم عرضه في إيطاليا أعلاه

تم تحليل البيانات الخاصة بإصابات فيروس كورونا في سوريا خلال تسلسل مرتب للبيانات يبدأ من تاريخ 1/22/2020 إلى تاريخ 7/14/2020 وقدرها 175يوم



ملاحظة للنموذج :القيم في ال20 يوم الاولى ليست 0 ولكن لصعوبة عرض المخطط لكافة القيم المتغيرة المتعلقة بالإصابات لكثرة عددها يكتفي بعرض التباين العالي بين القيم .

فكما نلاحظ في النموذج فقد بين لنا بداية ارتفاع تدريجي في عدد الإصابات في أول 40 يوم ليتحول بعدها إنتشار الفيروس إلى نمو اسي دون الإستقرار وعليه يمكننا الإستنتاج بأن عدد الإصابات في سوريا لم تصل بعد إلى الحد أي انها مستمرة وستستمر بالأرتفاع .

وعليه كما هو متبين ومقارنته بالوضع الحاضر في سوريا نستنتج أن النموذج سليم وهو يتناسب طرداً بما نلاحظه من زيادة كبيرة في عدد الإصابات كما يتم التصريح به من قبل منظمة الصحة في سوريا

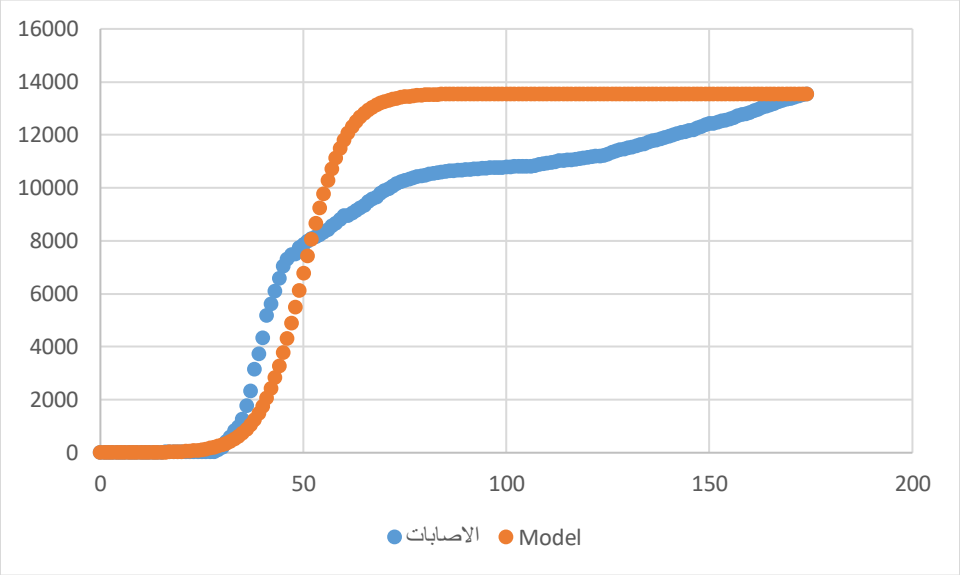
هذا يدل على أن الأوضاع لم تتحسن بل هي في إزدياد من وتيرة الإصابات .

South Korea / كوريا الجنوبية

سوف نستعرض النموذج اللوجستي الخاص ب كوريا الجنوبية وبسبب,

ما تداولته التصريحات لمؤسسات الصحية في كوريا فقد استطاعو من السيطرة على المرض . عبر الإجرائات الإحترازية , ليكون معيار نستطيع المقارنه به مع النموذج الذي نعمل عليه .

تم تحليل البيانات الخاصة بإصابات فيروس كورونا في ايطاليا خلال تسلسل مرتب للبيانات يبدأ من تاريخ 1/22/2020 إلى تاريخ 7/14/2020 وقدرها 175يوم



ملاحظة للنموذج: القيم في ال23 يوم الاولى ليست 0 ولكن لصعوبة عرض المخطط لكافة القيم المتغيرة المتعلقة بالإصابات لكثرة عددها يكتفي بعرض التباين العالي بين القيم .

فكما نلاحظ في النموذج فقد بين لنا بداية ارتفاع تدريجي في عدد الإصابات في أول 50 يوم ليتحول بعدها إنتشار الفيروس إلى التصاعد بشكل خفيف ثم ليحافظ على معدل مستقر من اليوم 80 إلى اليوم 127 ثم ليعاود الإرتفاع البطيئ إلى نمو

أما بالنسبة للمنحني اللوجستي فقد بين لنا وصول كوريا الجنوبية إلى حد الذروة في الإصابات عند 14000 اصابه عبر مقاطعة منحي اللوجستي مع منحني الإصابات

وعليه كما هو مبين ومقارنته بالوضع المصرح به في كوريا نستنتج أن النموذج سليم وهو يتناسب طردا بما نلاحظه إنحسار ووصول الإصابات إلى الذروة .

ملاحظة الباحث

لقد وجدنا عبر نموذج النوم اللوجستي امكانية جيدة للتطبيق على معدل الإصابات بفيروس كورونا لكل من كوريا الجنوبية وايطاليا وسوريا ولكن إن لاستخدام هذه المعلومات حقا في الحياة الواقعية ، سيكون من الضروري إجراء الكثير من التحقق من صحة النموذج ومقارنة الدقة ومقاييس الأداء الأخرى للنماذج المختلفة ومتابعة ما إذا كانت الاتجاهات المستقبلية تتبع النموذج المحدد.

يمكن على سبيل المثال استخدام هذا النوع من المعلومات من قبل صانعي السياسات لتقدير كيفية اتخاذ التدابير الصحيحة.

المقاربة الرابعة: إقتراح نموذج لتحسين جودة النتائج الطبية انطلاقاً من تجميع بيانات أعراض الإصابات وفق عناقيد لاكتشاف الأنماط المخفية

التطبيق العملي باستخدام : k means clustering algorithm

والسبب في استخدام الباحث للعنقدة على بيانات الأعراض المصاحبة لإصابات فيروس كورونا ..هو لقدرة خوارزمية K means على أكتشاف المعلومات المخفية ضمن الأنماط , لتكون المعلومات المستخرجة مفيدة قدر الإمكان من قبل صانعي السياسات لتقدير كيفية اتخاذ التدابير الصحيحة.

بداية تم تطبيق منهجية العنقدة من قبل الباحث باستخدام خوارزمية K means على مجموعة من البلدان التالية:

China
Italy
Iran
Republic of Korean
France
Spain
Germany
UAE
Other

استخدم الباحث في تطبيق الخوارزمية برنامج Weka المختص في التنقيب على البيانات , حيث كانت المدخلات التي أضافها الباحث إلى البرنامج وذلك لبناء النموذج المطلوب هي التالية :

1 – أدخل الباحث مجموعة البيانات المنظفة الخاصة بالأعراض المصاحبة لإصابات فيروس كورونا

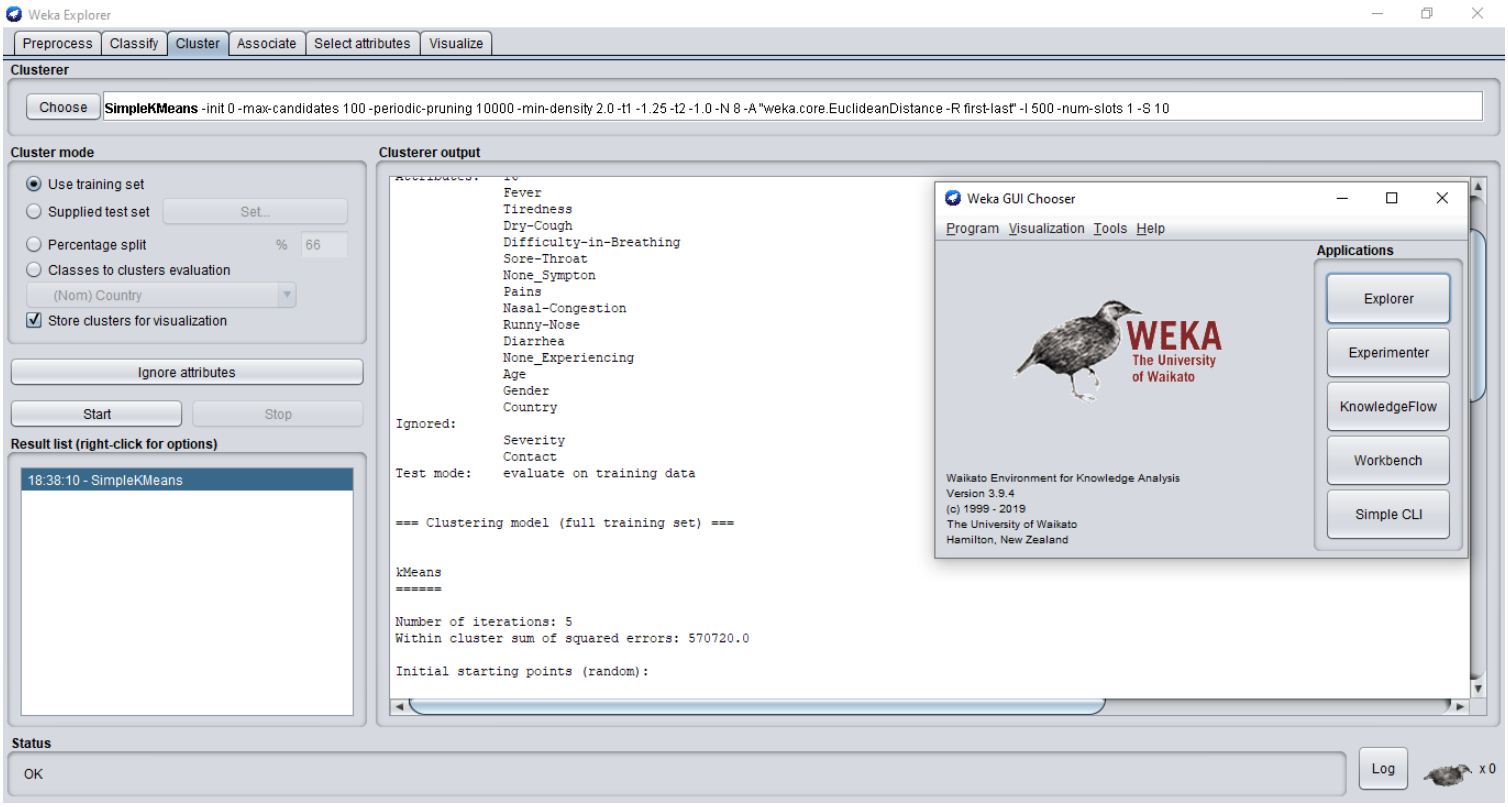
2– استخدم الباحث لحساب المسافات بين العقد المسافة الإقليدية **Euclidean distance**

3– تم تحديد عدد العناقيد K المراد عرضهم وتطبيق الخوارزمية عليهم في 8 ثمانية عناقيد

4– تم تحديد خوارزمية العمل بأنها خوارزمي K means

5– تم إدخال عدد أقصى من الدورات الحسابية Number of iterations وذلك لتمكين البرنامج من القدرة المطلوبة في حال كثرت الدورات .

خلصت النتائج إلى التالي:



Clusterer output

```
Missing values globally replaced with mean/mode

Final cluster centroids:
                                Cluster#
Attribute               Full Data          0          1          2          3          4          5          6          7
                       (140800.0) (35008.0) (14304.0) (28368.0) (20856.0) (13552.0)  (9464.0) (14240.0)  (5008.0)
=============================================================================================================================
Fever                          No         No         No         No         No         No         No         No         No
Tiredness                     Yes        Yes         No         No        Yes        Yes         No         No         No
Dry-Cough                     Yes        Yes         No         No        Yes        Yes         No         No         No
Difficulty-in-Breathing       Yes        Yes         No         No         No         No         No        Yes         No
Sore-Throat                    No        Yes        Yes         No         No         No         No         No        Yes
None_Sympton                   No         No         No         No         No         No         No         No         No
Pains                          No         No         No        Yes         No         No         No         No         No
Nasal-Congestion              Yes        Yes         No        Yes         No         No         No        Yes         No
Runny-Nose                    Yes         No        Yes         No        Yes        Yes         No        Yes         No
Diarrhea                       No         No        Yes         No        Yes        Yes         No         No        Yes
None_Experiencing              No         No         No         No         No         No        Yes         No         No
Age                     0 to 9  10 th 19  25 to 59  25 to 59       60+  25 to 59  20 to 24  10 th 19    0 to 9
Gender                    male     Female       male       male      male     Female       male     Female     Female
Country                  China      Spain  Other-EUR      China  Other-EUR      Other      Italy     France    Germany



Time taken to build model (full training data) : 3.98 seconds

=== Model and evaluation on training set ===

Clustered Instances
```

نسب تجمع السمات داخل العناقيد كانت على الشكل التالي:

العنقود	عدد الواصفات	نسبة التوزيع
0	35008	(25%)
1	14304	(10%)
2	28368	(20%)
3	20856	(15%)
4	13552	(10%)
5	9464	(7%)
6	14240	(10%)
7	5008	(4%)

الوقت المستغرق لبناء النموذج:

Time taken to build model (full training data) : 3.98 seconds

عدد الدورات التي أجراها البرنامج :

Number of iterations: 5

نتائج عملية العنقدة :

Attribute	Full Data	Cluster(0)	Cluster(1)	Cluster(2)	Cluster(3)	Cluster(4)	Cluster(5)	Cluster(6)	Cluster(7)
	(140800.0)	**(35008.0)**	**(14304.0)**	**(28368.0)**	**(20856.0)**	**(13552.0)**	**(9464.0)**	**(14240.0)**	**(5008.0)**
Fever الحمى	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**No**
Tiredness إرهاق	**Yes**	**Yes**	**No**	**No**	**Yes**	**Yes**	**No**	**No**	**No**
Dry-Cough سعال جاف	**Yes**	**Yes**	**No**	**No**	**Yes**	**Yes**	**No**	**No**	**No**
Difficulty-in-Breathing صعوبة في التنفس	**Yes**	**Yes**	**No**	**No**	**No**	**No**	**No**	**Yes**	**No**
Sore-Throat إلتهاب الحلق	**No**	**Yes**	**Yes**	**No**	**No**	**No**	**No**	**No**	**Yes**
None_Sympton / لا شيء من الاعراض الثانية	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**No**
Pains آلام	**No**	**No**	**No**	**Yes**	**No**	**No**	**No**	**No**	**No**
Nasal-Congestion / إحتقان بالأنف	**Yes**	**Yes**	**No**	**Yes**	**No**	**No**	**No**	**Yes**	**No**
Runny-Nose / سيلان الأنف	**Yes**	**No**	**Yes**	**No**	**Yes**	**Yes**	**No**	**Yes**	**No**
Diarrhea إسهال	**No**	**No**	**Yes**	**No**	**Yes**	**Yes**	**No**	**No**	**Yes**
None_Experiencing / لا شيء يعاني	**No**	**No**	**No**	**No**	**No**	**No**	**Yes**	**No**	**No**
Age العمر	**0 to 9**	**10 to 19**	**25 to 59**	**25 to 59**	**60+**	**25 to 59**	**20 to 24**	**10 to 19**	**0 to 9**
Gender الجنس	**male**	**Female**	**male**	**Male**	**male**	**Female**	**male**	**Female**	**Female**
Country البلد	**China**	**Spain**	**Other-EUR**	**China**	**Other-EUR**	**Other**	**Italy**	**France**	**Germany**

تفسير النتائج:

سنفسر النتائج المترتبة عن عملية العنقدة ووفق تجميع البيانات في **8** عناقيد على الشكل التالي :

- العنقود الاول : استطاعت الخوارزمية أن تبين لنا المعلومات التفصيلية التالية:

الشخص الذي جنسه انثى ويتراوح عمرها بين 10 و 19 عام , وهي مقيمة في اسبانيا قد تتجمع معها هذه الأعراض التالية (احتقان في الأنف – التهاب الحلق – صعوبة في التنفس – سعال جاف – ارهاق)

- العنقود الثاني : الشخص الذي جنسه ذكر , وعمره يتراوح بين 25 و 59 وهو(من بلاد اخرى) أي هو من خارج الصين و ايران و اميركا وايطاليا وكوريا وفرنسا واسبانيا والمانيا , قد تتجمع معه الأعراض التالية (إسهال – سيلان أنفي – آلام – إلتهاب الحلق)

- العنقود الثالث : الشخص الذي جنسه ذكر وعمره يتراوح بين 25 و 59 وهو من الصين , قدتتجمع معه الأعراض التالية (إحتقان بالأنف – آلام)

- العنقود الرابع :الشخص الذي جنسه ذكر وعمره يكون فوق ال60 سنة وهو من خارج ال(الصين و ايران و اميركا وايطاليا وكوريا وفرنسا واسبانيا والمانيا) قد تتجمع معه الأغراض التالية (إسهال – سيلان أنف – سعال جاف – إرهاق)

- العنقود الخامس : الشخص الذي جنسه انثى وعمرها بين 25 و 59 وهو من خارج (الصين و ايران و اميركا وايطاليا وكوريا وفرنسا واسبانيا والمانيا) قد تتجمع معها الأعراض التالية (إسهال – سيلان أنفي – سعال جاف)

- العنقود السادس : الشخص الذي جنسه ذكر وعمره يتراوح بين 20و24 وهو من ايطاليا قد تتجمع معه الأعراض التالية (لا شيئ اي قد يكون مصاب ولا يعاني من اعراض)

- العنقود السابع: الشخص الذي جنسه انثى وعمرها يتراوح بين 10 و 19 وهي من فرنسا قد تتجمع معها الأعراض التالية (سيلان أنف – أحتقان بالأنف – صعوبة في التنفس)

- العنقود الثامن : الشخص الذي جنسه انثى وعمره بين 0و9 وهي من المانيا قد تتجمع معها الأعراض التالية (إسهال – إلتهاب حلق)

من هذه النتائج المذكروة يمكن أستخدامها من قبل مسؤولي الرعاية الصحية لدعم جودة النتائج الطبية عن طريق توقع أن الشخص الذي يمتلك تلك الأعراض المذكورة أعلاه وتتناسب واصفاته من جنس وعمر ومدينة .. فقد يكون مصاب بفيروس كورونا الأمر الذي يستدعي إتخاذ تدابير وإجرائات مناسبة للحالة

ملاحظة الباحث :

الباحث ليس متخصص في الصحة أو أخصائيي في الأوبئة ، ولا ينبغي تفسير آراء هذا البحث على أنها نصيحة مهنية طبية بل هي مساعدة لتقدم نظرة من منظور رياضي أخر قد تكون مساعدة في عملية دعم جودة النتائج الطبية .

الفصل الرابع

النتائج والتوصيات و آفاق البحث المستقبلية

"I wanted to save the World"

ملخص نتائج البحث :

بناء على ما تم استعراضه خلال هذه الدراسة التي أجريت على البيانات الزمنية وبيانات الأعراض لفيروس كورونا , ومن خلال ما قدمه الباحث من إيضاح في الفصلين النظري والعملي , يمكن تلخيص أهم النتائج التي تم التوصل إليها بما يلي :

1– تم إثبات إمكانية مبدأية من التنبؤ بعدد الإصابات المتوقع حدوثها خلال يوم معين وقد كانت دقة النموذج في النتائج التي تم التوصل لها مقارنة بالنتائج التي نشهدها في الوقت الحاضر جيدة ويتفرع من هذه النتيجة ما يلي:

- استخدام تحليل الانحدار للتنبؤ بنتيجة ما خلال فترة زمنية معينة بناء على البيانات الزمنية السابقة يعد نموذج جيد ولكن لفترات تنبؤ محدودة
- استعرض النموذج المطبق على كل من (سوريا – كوريا الجنوبية – إيطاليا) بأن العلاقة بين عدد الإصابات وعدد الوفيات وعدد حالات الشفاء تتبع العلاقة إيجابية: إذ أظهر الخط اتجاها ً تصاعديا ً. وهذا يستوجب إلى أنه مع زيادة المتغير المستقل ، يزداد المتغير التابع أيض ًا.

2– ثم إثبات قدرة السلاسل الزمنية في التعامل مع بيانات الأوبئة حيث خلصت النتيجة إلى تمكن الباحث من التنبؤ بعدد الإصابات المتوقع حصولها خلال فترة زمنية قدرها شهر مقسمة الإصابات على توزع 4 أسابيع وذلك باستخدام طريقة المتوسط المتحرك على بيانات 5 أشهر السابقة , وبمقارنة النتائج التي حصل عليه الباحث من النموذج مقارونة بالنتائج الحقيقية على أرض الواقع للشهر السابع , وفق ما طرحته منظمة الصحة العالمية نجد أن النموذج كانت نتائجة التنبؤية خلال فترة شهر دقيقة بنسبة ٩٠ % وعليه يتفرع لدينا النتائج التالية :

- يمكننا التنبؤ بمعرفة كيف يتغير وينتشر الفيروس بمرور الوقت وما يصحبه من عدد الإصابات.
- تم إثبات أن معدل الإنتشار الخاص بفيروس كورونا عالمياً متجه باتجاه أسي المتزايد تدريجياً , وذلك عن طريق استخدام التحليل الاستكشافي لعرض مرئي للبيانات
- يمكن استخدام بيانات الأشهر الماضية و بيانات الشهر المتنبئ به لتوقع عدد الإصابات بعد فترة زمنية أطول
- كما بين النموذج بأن إنتشار الفيروس مذبذب هذا يؤدي إلى عدم تطابق واضح بين القيم الحقيقية والقيم التي تم التنبؤ بها يعطي بأن الفيروس لا يمكن التنبؤ بآلية انتشاره وإن الإنتشار لا يسير وفق آلية منتظمة

٣- تم إثبات إمكانية مبدئية بأن نموذج الانحدار اللوجستي يمكن أن يستخدم في نمذجة البيانات الوبائية للفترات المتقدمة ومنه نستنتج التالي :

- خلصت النتيجة الأولى للنموذج بأن ايطاليا قد وصلت إلى حد الذروة في عدد الإصابات وفق ما قدمه نموذج النمو اللوجستي .. وتم التأكد من صحة النموذج بالبيان الصادر عن الحكومة الأيطاليه بأنها قد وصلت بالفعل إلى حد الذروة
- خلصت النتيجة الثانية بأن سوريا لم تصل بعد إلى حد الذروة في عدد الإصابات وفق ما قدمة النموذج اللوجستي .. وبتأكيد على ذلك هو تصريحات وزارة الصحة بأننا في مرحلة الزيادة في عدد الإصابات
- خلصت النتيجة الثالية بأن كوريا الجنوبية قد وصلت تقريباً بشكل مبداي إلى مرحلة الذروة في عدد الإصابات
- النتيجة الرابعة بأن الفترات الأولى من الوباء يمكن نمذجتها بنموذج اسي أما في المراحل المتقدمة يجب الإستعانة بالنموذج اللوجستي

٤- توصل النتائج الخاصة بعملية التنقيب إلى بناء نموذج يمكن أن يقدم تقسيم للأشخاص المصابين بفيروس كورونا وارتباطهم بالأعراض المصاحبة للفيروس عن طريق الجنس والعمر والمدينة لنستنتج التالي :

- الشخص الذي جنسه انثى ويتراوح عمرها بين 10 و 19 عام , وهي مقيمة في إسبانيا

قد تتجمع معها هذه الأعراض التالية (احتقان في الأنف – التهاب الحلق – صعوبة في التنفس – سعال جاف – ارهاق)

- الشخص الذي جنسه ذكر , وعمره يتراوح بين 25 و 59 وهو(من بلاد اخرى) أي هو من خارج الصين و ايران و اميركا وإيطاليا وكوريا وفرنسا واسبانيا وألمانيا , قد تتجمع معه الأعراض التالية (إسهال – سيلان أنفي – آلام – إلتهاب الحلق)

- الشخص الذي جنسه ذكر وعمره يتراوح بين ٢٥ و ٥٩ وهو من الصين , قد تتجمع معه الأعراض التالية (احتقان بالأنف – آلام)

- الشخص الذي جنسه ذكر وعمره يكون فوق ال٦٠ سنة وهو من خارج ال(الصين و ايران و اميركا وايطاليا وكوريا وفرنسا واسبانيا والمانيا) قد تتجمع معه الأغراض التالية (إسهال – سيلان أنف – سعال جاف – إرهاق)

- الشخص الذي جنسه انثى وعمرها بين ٢٥ و ٥٩ وهو من خارج (الصين و ايران و اميركا وايطاليا وكوريا وفرنسا واسبانيا والمانيا) قد تتجمع معها الأعراض التالية (إسهال – سيلان أنفي – سعال جاف)

- الشخص الذي جنسه ذكر وعمره يتراوح بين ٢٠و٤٢ وهو من ايطاليا قد تتجمع معه الأعراض التالية (لا شئ ,أي قد يكون مصاب ولا يعاني من اعراض)

- الشخص الذي جنسه انثى وعمرها يتراوح بين ١٠ و ١٩ وهي من فرنسا قد تتجمع معها الأعراض التالية (سيلان أنف – أحتقان بالأنف – صعوبة في التنفس)

- الشخص الذي جنسه انثى وعمره بين ٥٠و٩٠ وهي من المانيا قد تتجمع معها الأعراض التالية (إسهال – إلتهاب حلق)

التوميات:

يمكن أستخدام النتائج السابقة التي توصل لها الباحث في المساعدة على دعم جودة النتائج الطبية , لتكون ورقة قوة إضافية من منظور رياضي آخر.

كما يوصي الباحث وفق ما توصل إليه من نتائج عبر إجراء العديد من النماذج الرياضية

بأن إنتشار فيروس كورونا في بلدنا سوريا مايزال في مراحل إنطلاقته بشكل اسي أي يمكن إحتواء المرض منذ بدايته عبر القليل من إجرائات الرعاية الصحية.

كما نوصي بوجوب إتباع العديد من هذه الأبحاث ودعمها لأنها تقدم منظور آخر من شأنه أن يحدث فرقاً في مجال الرعاية الصحية .. كما يوصى بتشكيل فرق بحوث علمية من كافة الإختصاصات للمساعدة على فهم هيكيلة هذا الفيروس .

- يمكن لمن له رغبة أن يتوسع في الدراسة القائمة وأن يستفيد من خوارزميات وطرق أخرى من شأنها أن تقدم منظور مختلف وقد يكون أفضل سعياً للوصول إلى نتائج أثمر

- إعطاء التنقيب في البيانات وتحليل البيانات أهمية أكبر داخل الشركات والمؤسسات البحثية لما تحمله من أدوات تساعد على معرفة نتائج ومعلومات كان من الصعب الوصول اليها سابقاً .

المراجع:

Charu C. Aggarwal / Data Mining

Steven S. Skiena / Data Science Design

Jiawei Han/ Data mining Concepts and Techniques

WHO / Coronavirus disease 2019 (COVID-19) / 2020

https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200314-sitrep-54-covid-19.pdf?sfvrsn=dcd46351_6

OXFORD / coronavirus-source-data

https://ourworldindata.org/coronavirus-source-data

informationisbeautiful/ visualizations/covid-19

https://informationisbeautiful.net/visualizations/covid-19-coronavirus-infographic-datapack/

semanticscholar/Discover New Insights About the Novel Coronavirus

https://www.semanticscholar.org/cord19

HDX / Data Responsibility for COVID-19

https://data.humdata.org/

worldometers/ COVID-19 CORONAVIRUS PANDEMIC

https://www.worldometers.info/coronavirus/#page-top

educba/ Data Science vs Data Mining

https://www.educba.com/data-science-vs-data-mining/

healthmap/ Confirmed cases worldwide

https://www.healthmap.org/covid-19/

data.world / dataset

https://data.world/resources/coronavirus/#jhd

MIT / about covid 19

https://www.covidanalytics.io/dataset

1 point3acres / Covid 19 map

https://coronavirus.1point3acres.com/en

The New England Journal of Medician /

https://www.nejm.org/coronavirus

wolframcloud/ Logistic-Growth-Model-for-COVID-190

https://www.wolframcloud.com/obj/covid-19/Published/Logistic-Growth-Model-for-COVID-19.nb

Data set

https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset

Data set

https://www.kaggle.com/roche-data-science-coalition/uncover

towardsdatascience / Getting Started With Weka 3 — Machine Learning on GUI

https://towardsdatascience.com/getting-started-with-weka-3-machine-learning-on-gui-7c58ab684513

towardsdatascience / What is Exploratory Data Analysis

https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15

towardsdatascience / Data Science for Beginners

https://towardsdatascience.com/data-science-for-beginners-850c3376a34a

towardsdatascience / A Road Map for Data Science

https://towardsdatascience.com/a-road-map-for-data-science-d1977504a72b

towardsdatascience / End to End Time Series Analysis and Modelling

https://towardsdatascience.com/end-to-end-time-series-analysis-and-modelling-8c34f09a3014

towardsdatascience / Almost Everything You Need to Know About Time Series

https://towardsdatascience.com/almost-everything-you-need-to-know-about-time-series-860241bdc578?source=false---------5

Towardsdatascience / Modeling Logistic Growth

https://towardsdatascience.com/modeling-logistic-growth-1367dc971de2

towardsdatascience