الجمهورية العربية السورية

وزارة التعليم العالي

معهد العالي لإدارة الأعمال

نظام توصية باستخدام التنقيب في المعطيات

حالة عملية على محل بيع حلويات

Recommendation System Using Data Mining

إعداد الطالبة

دارين عبد الحق

إشراف الدكتور

د. كادان الجمعة

2022-2021

مشروع أعد لنيل درجة البكالوريوس في إدارة الأعمال اختصاص إدارة العمليات والمعلومات

إهداء

إلى من علمني معنى الحياة ودفعني إلى النجاح... إلى مثال القوة والعنفوان والحنان

أمي أديلا

إلى من علمني الرصانة والاعتدال... إلى من آمن بي

أبي حازم

إلى من أثق بهم بحياتي ومن شاركوني عبء الحياة وقاسموني لحظات الفرح

أخوتي ياسمين وتولين

إلى من كان مصدر إلهامي... إلى المرأة التي أحلم أن أشبهها

جدتي فريال

إلى من شاركوني في مسيرتي الدراسية، ولهم كل الفضل في اجتياز أهم سنين حياتي

أصدقائي

كلمة شكر

أتوجه بخالص الشكر والتقدير لحضرة الدكتور المشرف كادان الجمعة، الذي بفضله قد اخترت اختصاصي وأتممت مسيرتي الدراسية.

كما أتوجّه ببالغ الشكر والتقدير لكافة أعضاء الكادر التدريسي والإداري لمعهد العالي لإدارة الأعمال من جهد ومعرفة.

دارين عبد الحق

ملخص البحث

التنقيب في المعطيات هو حقل متعدد التخصصات، يستفيد من دراسة العديد من المجالات بما في ذلك تقنية قاعدة البيانات، الذكاء الاصطناعي، والتعلم الآلي، والتعرف على الأنماط، وإلخ. فتعد مهمة التنقيب في المعطيات هي إيجاد واستخلاص معلومات محددة من كميات كبيرة من البيانات بصورة عامة. في تطبيق نظام التوصية، يعد استخراج البيانات تقنية تستخدم للعثور على قواعد التوصية. ويتألف من عدة مهام مثل التصنيف، والتجميع، والتنقيب عن قواعد الارتباط، والتلخيص، إلخ. يهدف البحث الى تصميم نظام توصية يساعد الزبائن والمستخدمين لاختيار أنسب المواد التي تلائم رغباتهم، وبذلك زيادة من أرباح المحل في المستقبل. ومحاولة التوصل إلى توصيات لتحسين من النظام التوصية المقترح للمستقبل.

وقد طبّقت خوارزمية Apriori على مجموعة بيانات الخاصة بمحل مبيع الحلويات Cookie Monster، وهي معاملات الحاصلة عن طريق تطبيق توصيل طلبات الطعام Beeorder. تم شرح الخوارزمية ومن ثم كيفية تطبيقها، وتجربة بعض الاحتمالات الواقعية على النظام التوصية هذا.

كانت أهم توصيات البحث هي تحسين من النظام التوصية من خلال الحصول على بيانات أكثر لتحسين من دقة النتائج، وإضافة مقاييس أخرى كتقييم الزبائن للمنتجات للتعرف على رغباتهم بشكل أدق وأحسن.

الكلمات المفتاحية: *التنقيب في المعطيات، خوارزمية أبريوري، نظم التوصية*

Abstract

Data mining is an interdisciplinary field, benefiting from the study of many areas including database technology, artificial intelligence, machine learning, pattern recognition, etc. The task of data mining is to find and extract specific information from large amounts of data in general. In recommender application, data mining is a technique used to find recommender rules. It consists of several tasks such as classification, grouping, association rule mining, summarization, etc. The research aims to design a recommendation system that helps customers and users choose the most appropriate materials that suit their desires, and thus increase the profits of the store in the future. And try to come up with recommendations for improvement of the proposed recommending system for the future.

Apriori's algorithm was applied to a dataset of Cookie Monster, dataset of transactions made by food delivery app Beeorder. The algorithm is explained and how to apply it, and some realistic possibilities are tested on this recommender system.

The most important recommendations of the research were to improve the recommendation system by obtaining more data to improve the

accuracy of the results, and adding other metrics such as customer evaluation of products to know their desires more accurately.

***Keywords**: Data Mining, Apriori Algorithm, Recommendation Systems*

الفهرس

قائمة الأشكال

الفصل الأول

الإطار التمهيدي

1- مقدمة

ظهرت في عصرنا الحالي طرقا جديدة للتواصل وتبادل المعلومات بسرعات عالية، حجوم كبيرة، وتنوع غير مسبوق؛ مما أدى إلى التخلص من الحدود الزمانية والمكانية. ساهمت شبكة الانترنت بربط الأشخاص ببعضهم ونشر المعلومات الشخصية والعامة في جميع أنحاء العالم بسهولة. هذا التطور جلب معه سبل جديدة التي قد تحسن من أرباح الشركات عبر التعرف على العملاء بشكل شخصي أكثر، أو بكلمات أبسط؛ تجميع بياناتهم.

تعد أنظمة التوصية جزءًا مهماً من نظام المعلومات والتجارة الإلكترونية. وأدى ما يقرب من عقدين من البحث حول النظم التوصية إلى مجموعة متنوعة من الخوارزميات ومجموعة غنية من الأدوات لتقييم أدائها. يعد التنقيب في المعطيات أحد أهم الطرق المستخدمة لاستخراج البيانات غير البديهية من قاعدة معطيات وتفسيرها للاستفادة منها لاحقا في تحقيق نتائج وأهداف. كما أنّه تستخدم خوارزمياته المختلفة في أنواع مختلفة من نظم التوصية. وفي حالتنا فإن خوارزمية Apriori هي التي ستستخدم في نظام التوصية.

2- دراسات سابقة

❖ دراسة (Pradana et al. 2019) بعنوان

Product Recommendation Systems using Apriori in the Selection of Shoe based on Android.

ملخص الدراسة: استخدمت هذه الدراسة نموذج الشلال بداخله أربع مراحل. تم اختبار النظام باستخدام طريقة Apriori لإنتاج قواعد مشتقة من نمط الجمع بين عنصرين، والقواعد واللواحق. نتائج هذه الدراسة هي البحث الناجح عن مستوى الدعم والثقة في الأحذية بحيث تكون الأحذية التي تم الحصول عليها أكثر دقة باستخدام التنقيب عن البيانات. من نتائج الحساب التي تم الحصول عليها أن أعلى منتج لقاعدة الارتباط هو New Balance> Adidas بقيمة 87.5٪ وأقل نتيجة على Adidas> Puma بقيمة 18.2٪.

❖ دراسة (Fatoni et al. 2018) بعنوان

Online Store Product Recommendation System Uses Apriori Method.

ملخص الدراسة: حلل هذا البحث القواعد في البيانات التاريخية للشراء من زوار المتجر عبر الإنترنت للحصول على توصية بالمنتجات التي سيتم عرضها. وتم استخدام قواعد الارتباط وخوارزمية Apriori. يتمثل تطبيق طريقة Apriori في هذا البحث في العثور على معظم مجموعة العناصر بناءً على بيانات المعاملة ثم تكوين نمط اقتران مجموعة العناصر. وفقًا لنتيجة التجربة، فإن قاعدة الارتباط قادرة على إنشاء توصيات دقيقة بقيم ثقة 76.92٪.

❖ دراسة نور نصّار 2019 بعنوان: نظام توصية ضمن معطيات كبيرة

ملخص الدراسة: قدمت في هذه الأطروحة نموذجًا جديدًا لنظام توصية متعدد المعايير يعتمد على الترشيح التعاوني ويستخدم التعلم العميق. يتألف النموذج من جزأين: في الجزء الأول، يحصل النموذج على خصائص المستخدمين وخصائص العناصر لاستخدامها كدخل لشبكة عصبونية عميقة تتوقّع التقييمات المتعددةُ. تشكل هذه التقييمات المتعددة الدخل للجزء الثاني، وهو شبكة عصبونيةُ عميقة تستخدم لتوقّع التقييم الكلّي.

أظهرت التجارب التي أجريت تفوق النموذج المقترح على طرق التوصية التقليدية في جميع المقاييس المستخدمة لتقييم الأداء. اقترح أيضاً نموذج ثاني محسن يعتمد على دمج الشبكة العصبونية العميقة مع طريقة تحليل المصفوفة إلى عوامل من
أجل توقّع التقييمات المتعددة. بيّنت التجارب تحسينا في أداء هذا النموذجً مقارنة بأداء النموذج الأول وباقي طرق التوصية التقليدية

❖ **دراسة (Singh et al. 2021) بعنوان**

Optimized recommendations by user profiling using apriori algorithm

ملخص الدراسة: اقترحت هذه الورقة طريقة جديدة وحديثة للتوصية المحسنة باستخدام مجموعة بيانات عدسة الفيلم Movie-lens. وهي تتألف من ملفات تعريف المستخدمين بناءً على تعليقاتهم على أفلام مختلفة في سيناريو مختلف. مجموعات البيانات التي تم جمعها باستخدام 19 نوعًا مقسمة إلى فئتين للإعجاب والكراهية بناءً على التصنيف المحدد. ثم تم تطبيق خوارزمية apriori للعثور على قائمة التفضيلات النهائية للمستخدمين. بعد ذلك، يتم استخدام هذه الميزات في مقاييس التشابه الحوسبية وأساليب التنبؤ بالتصنيف. توفر نظام التوصية المقترح المستند إلى CF دقة استرجاع محسّنة وقياس F1 ودقة أكثر من خوارزميات CF التقليدية الأخرى باستخدام تشابه Jaccard ومسافة جيب التمام وجميع المقاييس الأخرى. تعمل جميع مقاييس التشابه التقليدية على زيادة دقة التنبؤ بالتصنيف باستخدام نهجنا المقترح في ظل السيناريوهات المتفرقة المختلفة.

3- مشكلة البحث

تتجلى مشكلة البحث باختيار المستخدمين أو العملاء لعناصر من بين كم هائل من للعناصر البديلة. وبالتالي ممكن تلخيص مشكلة البحث بالسؤال التالي:

– كيف يمكن لنظم التوصية مساعدة الزبائن باختيار أمثل عنصر من بين مجموعة العناصر؟

4- أهداف البحث

يهدف البحث الى تصميم نظام توصية يساعد الزبائن والمستخدمين لاختيار أنسب المواد التي تلائم رغباتهم.

– بناء نظام توصية يوصي المتعاملين معه بالمواد التي ستشبع من رغباتهم

5- أهمية البحث

الأهمية النظرية: تتجلى من خلال توضيح العديد من المفاهيم والتعاريف المتعلقة بالتنقيب في المعطيات ونظم التوصية وكيفية بناء نظام توصية معتمد على الخوارزمية المتبعة وبالتالي يمكن أن يكون هذا البحث مرجعا متواضعا للمهتمين في هذا المجال.

الأهمية العملية: تتجلى من خلال تطبيق نظام التوصية هذا في المطعم والاستفادة من النتائج من خلال فهم العميل والتوصل إلى أكبر نسبة من الرضا لدى العميل

6- الحدود المكانية والزمانية

الحدود المكانية: مطعم حلويات المنزلية Cookie Monster

الحدود الزمانية: جمعت المعلومات، وهي المعاملات، من أول الشهر العاشر إلى أخر الشهر 5. مجموعها 195 معاملة.

المحددات: قلة المعاملات المتحصلة والذي يسبب بتغير دقة النتائج.

الفصل الثاني

القسم النظري

القسم الأول

التنقيب في المعطيات

تنقيب في المعطيات

التنقيب في المعطيات، والذي يشار إليه أيضًا باسم اكتشاف المعرفة في قواعد البيانات، هو عملية استخراج غير بديهية من المعلومات الضمنية، غير المعروفة سابقًا والتي يحتمل أن تكون مفيدة (مثل قواعد المعرفة والقيود والانتظام) من البيانات في قواعد البيانات.-Piatetsky) Shapiro et al, 1991)[1]

فهو يحلل كميات كبيرة من البيانات من أجل اكتشاف أنماط ذات مغزى وفائدة.

وفي تعاريف أخرى فإن التنقيب في المطيات هي عملية اكتشاف المعرفة الشيقة من كميات كبيرة من البيانات المخزنة في قواعد البيانات والبيانات المستودعات، أو مستودعات المعلومات الأخرى (Han et al., 2011).[2]

التنقيب عن المعطيات له قابلية تطبيق واسعة، مع تطبيقات في الذكاء والأمن التحليل وعلم الوراثة والعلوم الاجتماعية والطبيعية والأعمال. ومن الممكن دراسة الطبيعة الاستهلاكية لدى الزبائن أو مدى استجابتهم للإعلانات، أو نسبة أن العميل يرجح إلى الاحتيال أو التخلي عن خدمات الاشتراك ذات أهمية حيوية للأعمال.

[1] 1- G. Piatetsky-Shapiro and W.J. Frawley. Knowledge Discovery in Databases. 1991.
[2] 2- Jiawe Han, Micheline Kamber, Jian Pei. Data mining: Concepts and Techniques ,3rd ed. (Morgan Kaufmann publisher,2011).

بينما تركز تطبيقات الإحصاء التقليدية على مجموعات البيانات الصغيرة نسبيًا يتضمن تعدين كميات كبيرة جدًا وأحيانًا هائلة من المعلومات. نتحدث هنا عن ميغا بايت وتيرابايت من المعلومات.(Johannes Ledotler. 2013)[3]

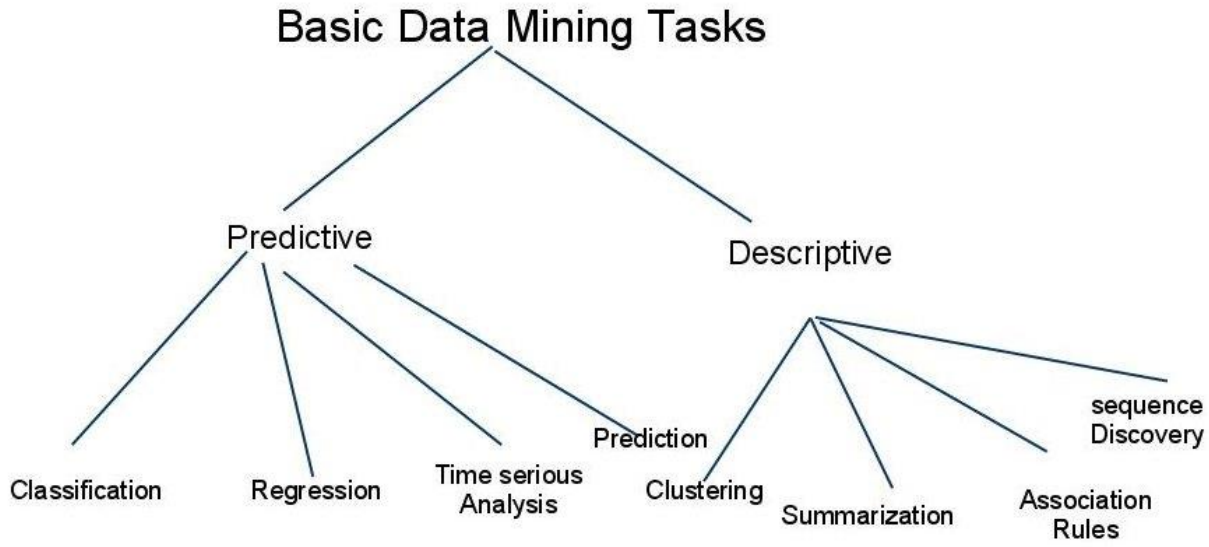
أنواع التنقيب في المعطيات

تخدم كل من تقنيات التنقيب في المعطيات التالية العديد من مشاكل العمل المختلفة وتوفر نظرة ثاقبة مختلفة لكل منها. ومع ذلك، فإن فهم نوع مشكلة العمل التي تحتاج إلى حلها سيساعد أيضًا في معرفة التقنية الأفضل للاستخدام، والتي ستحقق أفضل النتائج. يمكن تقسيم أنواع التنقيب عن البيانات إلى جزأين أساسيين على النحو التالي:

1- التنقيب في المعطيات التنبؤية
2- التنقيب في المعطيات الوصفية

يمكن تسمية عملية الاكتشاف كتنقيب عن البيانات الوصفية، في حين أن تصنيف مجموعة بيانات الاختبار باستخدام العلاقات المكتشفة ينظر إليها على أنها تنقيب عن البيانات التنبؤية. (Tiwari et al, 2009)[4]

[3] 4- Data Mining and Business Analytics with R, First Edition. Johannes Ledolter.(2013 John Wiley & Sons, Inc.)
[4] 3- CHOUDHARY, A.K., HARDING, J.A. and TIWARI, M.K., 2009. Data mining in manufacturing: a review based on the kind of knowledge



أنواع التنقيب في البيانات.1 رسم توضيحي

1-التنقيب في المعطيات التنبؤية: للتنبؤ بقيمة سمة معينة بناءً على قيم السمات الأخرى. يمكن أيضًا تقسيم التنقيب في المعطيات التنبؤية إلى أربعة أنواع مذكورة أدناه:

- تحليل التصنيف(Classification)
- تحليل الانحدار(Regression)
- تحليل السلاسل الزمنية(Time Series)
- تحليل التنبؤ(Prediction)

مثال عن تطبيق التنقيب في المعطيات التنبؤية: توقع تغيرات الأسعار

1-1 تحليل التصنيف (Classification) [5](Han et al, 2011)

تصنيف البيانات هو عملية تتكون من خطوتين، خطوة التعلم (حيث يتم إنشاء نموذج التصنيف) وخطوة التصنيف (حيث يستخدم النموذج للتنبؤ بتسميات الفئات لبيانات معينة).

في الخطوة الأولى، يتم إنشاء المصنف الذي يصف مجموعة محددة مسبقًا من فئات البيانات أو المفاهيم. هذه هي خطوة التعلم (أو مرحلة التدريب)، حيث يتم تصنيف خوارزمية تبني المصنف عن طريق تحليل أو "التعلم من" مجموعة تدريب مكونة من مجموعات قاعدة بيانات وتسميات الفئات المرتبطة بها. في سياق التصنيف، يمكن أن تكون مجموعات البيانات يشار إليها على أنها عينات أو أمثلة أو مثيلات أو نقاط بيانات أو كائنات.

في الخطوة الثانية، يتم استخدام النموذج للتصنيف. أولاً، يتم تقدير الدقة التنبؤية للمصنف. دقة المصنف في مجموعة اختبار معينة هي النسبة المئوية لمجموعة مجموعات الاختبار التي تصنف بشكل صحيح من قبل المصنف. تتم مقارنة تسمية الفئة المرتبطة بكل مجموعة اختبار مع توقع فئة المصنف الذي تم تعلمه لهذه المجموعة. إذا تم النظر في دقة المصنف على أنّه مقبول، يمكن استخدام المصنف لتصنيف مجموعات البيانات المستقبلية التي لا يعرف تصنيف الفئة لها.

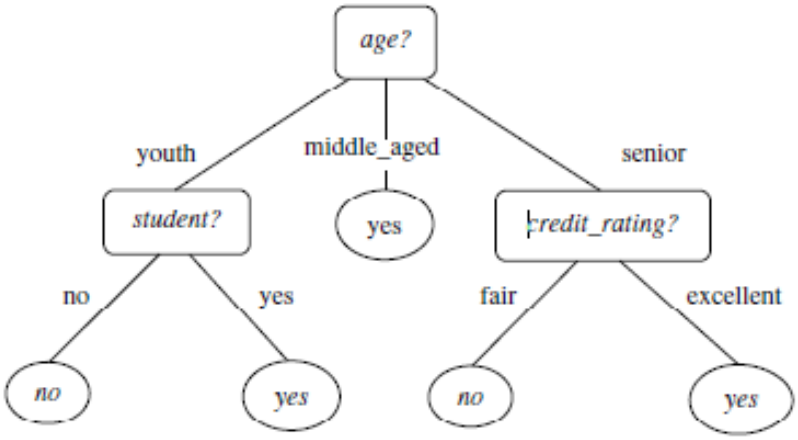
[5] 2- Jiawe Han, Micheline Kamber, Jian Pei. Data mining: Concepts and Techniques ,3rd ed. (Morgan Kaufmann publisher,2011).

1-1-1 خوارزميات تحليل التصنيف

أولا: استقراء شجرة القرار (Decision Tree Induction) (Han et al, 2011)[6]

استقراء شجرة القرار هو تعلم أشجار القرار من التدريب المسمى بمجموعات الفصول الدراسية. شجرة القرار عبارة عن هيكل شجري يشبه المخطط الانسيابي، حيث تكون كل عقدة داخلية (Non-leaf node) اختبار على سمة، كل فرع يمثل نتيجة اختبار، وتحمل كل عقدة ورقية (Terminal node) تسمية فئة. العقدة العليا في الشجرة هي عقدة الجذر.

يتم عرض شجرة القرار النموذجية في الشكل 2.



رسم توضيحي2: شجرة القرار النموذجية

[6] 2- Jiawe Han, Micheline Kamber, Jian Pei. Data mining: Concepts and Techniques ,3rd ed. (Morgan Kaufmann publisher,2011).

شجرة القرار الخاصة بالمفهوم شراء كمبيوتر، تشير إلى ما إذا كان العميل من المرجح أن يشتري جهاز كمبيوتر أم لا. تمثل كل عقدة داخلية (non-leaf) اختبارًا على سمة(Attribute) . تمثل كل عقدة طرفية فئة (إما تشتري كمبيوتر = نعم أو تشتري كمبيوتر = لا).

ثانيا: طرق التصنيف البايزية (Bayes Classification Methods)

المصنفات البايزية هي مصنفات إحصائية تستطيع توقع احتمالات عضوية الفئة مثل احتمال انتماء مجموعة معينة لفئة معينة. [7](Han et al, 2011)

تفترض مصنفات البايزية الساذجة(Naïve Bayes Classifiers) أن تأثير قيمة السمة على فئة معينة مستقلة عن قيم السمات الأخرى. هذا افتراض يسمى الاستقلال الشرطي الطبقي. وهي مصنوعة لتبسيط الحساب المتضمن، وبهذا المعنى، يعتبر "ساذج".[8](Ming Leung, 2007)

يعمل Naïve Bayes جيدًا في حالات متغيرات الإدخال الفئوية مقارنة بالمتغيرات العددية. إنه مفيد لعمل تنبؤات وتنبؤ بالبيانات بناءً على النتائج التاريخية.

[7] 2- Jiawe Han, Micheline Kamber, Jian Pei. Data mining: Concepts and Techniques ,3rd ed. (Morgan Kaufmann publisher,2011).
[8] 6- K. Ming Leung. Naïve Bayesian Classifier, 2007

ثالثا: التصنيف القائم على قواعد (Rule-Based Classifier) (Xiao Lili et al, 2014)[9]

يعتبر التصنيف القائم على قواعد، أو ما يسمى أيضا بتعلم القواعد، قيم بسبب المزايا التالية:

1. القواعد طبيعية جدًا لتمثيل المعرفة، كما يفهمها الناس وتفسيرها سهل.

2. نتائج التصنيف سهلة الشرح. بناء على قاعدة بيانات قاعدة البيانات وإدخال البيانات من المستخدم، يمكننا شرح القاعدة أو مجموعة القواعد المستخدمة لاستنتاج تسمية الفئة حتى يكون المستخدم واضحًا بشأن المنطق الكامن وراء الاستدلال.

3. يمكن بسهولة تحسين نماذج التصنيف المستندة إلى القواعد واستكمالها عن طريق إضافة قواعد جديدة من خبراء المجال. تم تنفيذ تلك الإضافات والتحسينات بنجاح في العديد من الأنظمة الخبيرة.

4. بمجرد تعلم القواعد وتخزينها في قاعدة بيانات القاعدة، يمكننا لاحقًا استخدامها لتصنيف حالات جديدة بسرعة من خلال بناء هياكل الفهرس للقواعد والبحث عن القواعد ذات الصلة بكفاءة.

[9] 7- Xiao-Li Li and Bing Liu. Rule-Based Classification, 2014

5. أنظمة التصنيف القائمة على القواعد قادرة على المنافسة مع خوارزميات التصنيف الأخرى وفي كثير من الحالات تكون أفضل منهم.

يتم تمثيل القواعد في شكل المنطق كعبارات IF-THEN، على سبيل المثال يمكن التعبير عن قاعدة شائعة الاستخدام على النحو التالي:

If *condition* **Then** *conclusion*

حيث يسمى الجزء IF بـ "سابقة" أو "حالة" ويكون الجزء Then يسمى "اللاحقة" أو "الاستنتاج". وهو يعني في الأساس أنه إذا كان شرط القاعدة ملبى، يمكننا أن نستنتج أو نخصم النتيجة. (Xiao Li Li et al, 2014)[10]

[10] 7- Xiao-Li Li and Bing Liu. Rule-Based Classification, 2014

2-1 تحليل الانحدار (Regression Analysis)

الانحدار هو معادلة تمثل الطريقة التي تشرح فيها مجموعة من العوامل النتيجة وكيف يتحرك الناتج مع كل عامل. لكن نموذج الانحدار لا يوضح سبب تحرك المتغيرات مع بعضها البعض، وبالتالي لم نقم بتوضيح سبب وجود تأثير سببي أو حجم أي تأثير سببي بسبب وجود تفسيرات بديلة لسبب تتحرك المتغيرات معًا. [11](Arkes et al, 2019)

الأهداف الأربعة الرئيسية لتحليل الانحدار: (Arkes et al, 2019)

الاستخدام الأكثر شيوعًا لتحليل الانحدار هو تحديد كيفية تأثير عامل سببيًا على الآخر. على سبيل المثال، إذا حصل شخص ما على سنة دراسية إضافية واحدة، فكم نتوقع ذلك زيادة دخل الفرد؟

ثانيًا، يمكن استخدام الانحدارات للتنبؤ أو التنبؤ بنتيجة. قد يرغب الجيش في الحصول على توقع جيد لعدد الجنود في السنة الأولى الذين سيتركون الخدمة حتى يتمكن الجيش من تحديد أهداف التجنيد المثلى لسنة معينة. الاستخدام الثالث للانحدارات هو تحديد تنبؤات بعض العوامل. ما هو أفضل تنبؤ / توقع للنتيجة؟

[11] 8- Jeremy Arkes. Regression Analysis: A Practical Introduction, 2019.

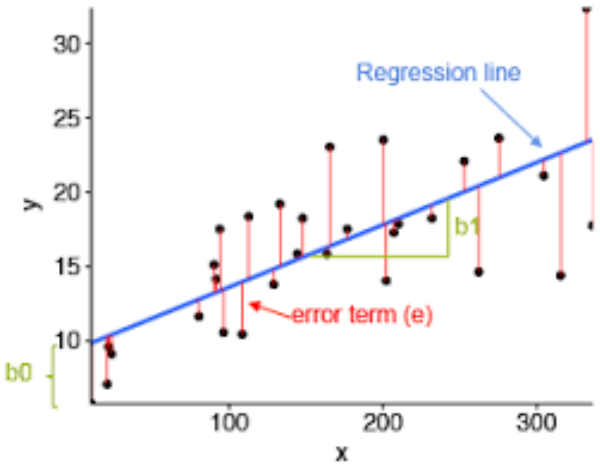
والاستخدام الرئيسي الرابع لتحليل الانحدار هو تعديل النتيجة لعوامل مختلفة. على سبيل المثال، بدلاً من مجرد تقييم فعالية المعلم بناءً على درجات اختبار طلابه، يمكننا تعديل هذه الدرجات بناءً على النتائج السابقة للطلاب وربما بناءً على التركيبة السكانية.

1-2-1 خوارزميات تحليل الانحدار:

أبسط شكل من أشكال الانحدار، الانحدار الخطي البسيط (Simple Linear Regression)، يستخدم صيغة خط مستقيم (y = a + bx).

حيث:

- x هو متغير المحور الأفقي.
- y هو متغير المحور الرأسي.
- a هو تقاطع y
- b هو منحدر الخط.



رسم توضيحي3 : بيان انحدار خطي بسيط

ويحدد قيمتي a وb المناسبة للتنبؤ بقيمة y بناءً على قيمة معينة لـ x. [12] (Gupta. 2015)

[12] 9- Swati Gupta. A Regression Modeling Technique on Data Mining, 2015

يستخدم الانحدار الخطي لتقدير العلاقة بين متغيرين. ويمكن التعبير عن العلاقة ب: ($y=a_0 + a_1x + e$) حيث e هي نسبة الخطأ، والتي تشير إلى مدى بُعد نقطة بيانات فردية عموديًا عن خط الانحدار الحقيقي. (Arkes. 2019)[13]

وأما في حالة الانحدار الخطي المتعدد (Multiple Linear Regression)، تصبح العلاقة بالشكل التالي: $y = a_0 + a_1x_1 + \cdots + a_m x_m + e$.

يتنبأ الانحدار الخطي المتعدد بوجود علاقة بين متغيرات متعددة. على سبيل المثال، هل هناك علاقة بين الدخل والتعليم وأين يختار المرء العيش؟

انحدار لاسو (Least Absolute Shrinkage and Selection Operator-Lasso Regression)وهو نوع من الانحدار الخطي يستخدم الانكماش. الانكماش هو المكان الذي يتم فيه تقليص قيم البيانات باتجاه نقطة مركزية، مثل المتوسط. يشجع إجراء lasso النماذج البسيطة والمتفرقة (أي النماذج ذات المعلمات الأقل). هذا النوع المعين من الانحدار مناسب تمامًا للنماذج التي تعرض مستويات عالية من تعدد الخطية (Multicollinearity) أو عندما تريد أتمتة أجزاء معينة من اختيار النموذج، مثل التحديد المتغير أو حذف المعلمة. (Beyer. 2002)[14]

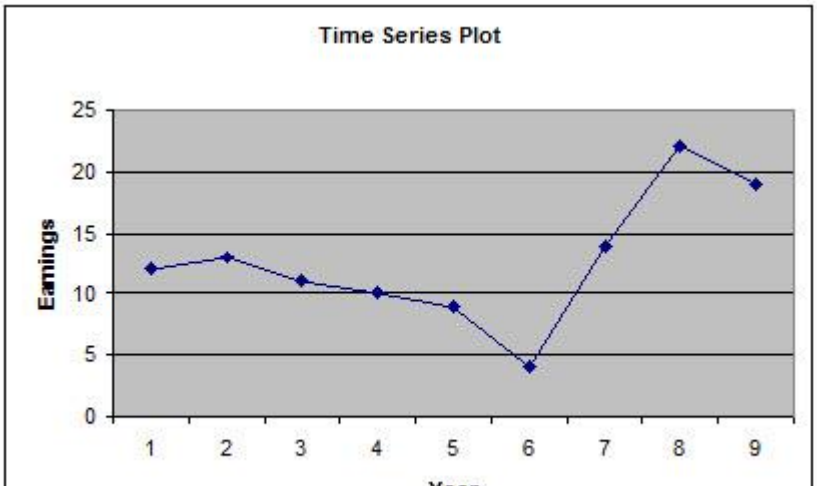
[13] 8- Jeremy Arkes. Regression Analysis: A Practical Introduction, 2019
[14] 10- Beyer, W. H. CRC Standard Mathematical Tables and Formulae, 31st ed. 2002

3-1 تحليل السلاسل الزمنية (Time Series Analysis)

السلاسل الزمنية هي سلسلة من الملاحظات المأخوذة على فترات منتظمة من الوقت مثل قراءات درجة الحرارة بالساعة وأسعار الأسهم اليومية. [15](Pena et al, 2001) وتعرف أيضا بأنها مجموعة من الملاحظات يتم إجراؤها بالتتابع في الوقت المناسب. (Chakarvert et al. 2019)[16]

ومن أمثلتها التي تنشأ في الممارسة:

1- السلاسل الزمنية الاقتصادية كأسعار الأسهم في الأيام المتتالية ومجموع الصادرات في الأشهر المتتالية وأرباح الشركة في السنوات المتتالية.



رسم توضيحي4 : بيان سلسلة زمنية اقتصادية

[15] 11- Daniel Peña, George C. Tiao, Ruey S. Tsay. A Course in Time Series Analysis. 2001

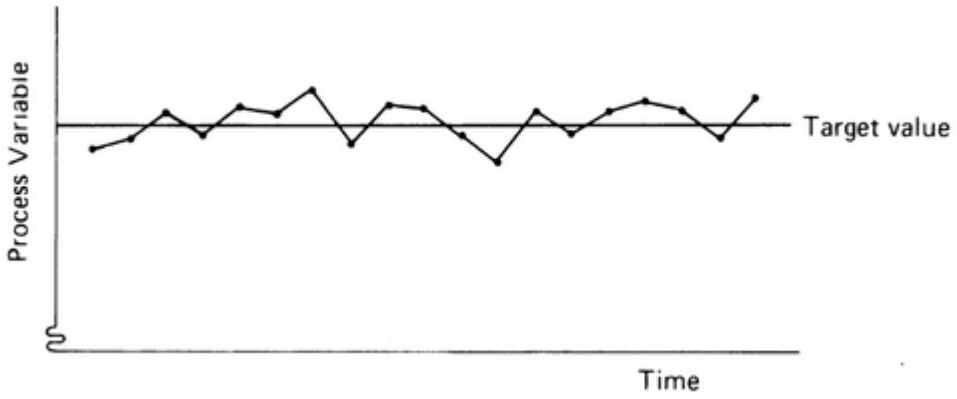
[16] 12- Mohini Chakarverti, Nikhil Sharma and Rajiva Ranjan Divivedi. Prediction Analysis Techniques of Data Mining: A Review. 2019

2-السلاسل الزمنية الفيزيائية تستخدم أنواع عديدة من السلاسل الزمنية في العلوم الفيزيائية، وخاصة في علم الأرصاد الجوية وعلوم البحار والجيوفيزياء.

3-السلاسل الزمنية التسويقية تعتبر أرقام المبيعات التحليلية في الأسابيع أو الأشهر المتتالية مشكلة مهمة في التجارة.

4-السلاسل الزمنية الديموغرافية يريد علماء الديموغرافيا توقع التغيرات في عدد السكان لما يصل إلى 10 إلى 20 عامًا في المستقبل.

5-تحكم العملية في التحكم في العملية، تكمن المشكلة في اكتشاف التغيرات في أداء عملية التصنيع عن طريق قياس متغير يوضح جودة العملية. يمكن رسم هذه القياسات مقابل الوقت كما في الشكل 5. عندما تبتعد القياسات كثيرًا عن الهدف، يجب اتخاذ الإجراءات



رسم توضيحي5 : بيان تحكم العملية

التصحيحية المناسبة للتحكم في العملية.

6-العمليات الثنائية تحدث السلاسل الزمنية من هذا النوع بشكل خاص في نظرية الاتصال، حيث يمكن أن تأخذ الملاحظات واحدة من قيمتين فقط، يُشار إليها عادةً بالرقم 0 و 1.

7- عمليات النقطة بالنسبة للسلاسل من هذا النوع، نحن مهتمون بتوزيع عدد الأحداث التي تحدث في فترة زمنية معينة وكذلك في توزيع الفترات الزمنية بين الأحداث. نستخدم هذا النوع عندما نعتبر بأن الأحداث المراقبة عشوائية.

الأهداف الرئيسية لتحليل السلاسل الزمنية هي: (Pena et al. 2001)[17]

1-فهم البنية الديناميكية أو المعتمدة على الوقت لملاحظة سلسلة واحدة – تحليل السلاسل الزمنية بمتغير واحد(Univariate) . سيساعد معرفة الهيكل الديناميكي في إنتاج تنبؤات دقيقة للملاحظات المستقبلية وتصميم مخططات التحكم المثلى.

2- التأكد من العلاقات الرائدة والمتأخرة والتغذية الراجعة بين عدة سلاسل – تحليل السلاسل الزمنية متعددة المتغيرات(Multivariate) .

[17] 11- Daniel Peña, George C. Tiao, Ruey S. Tsay. A Course in Time Series Analysis. 2001

4-1 تحليل التنبؤ (Prediction Analysis)

تحليل التنبؤ هو طريقة عامة للتنبؤ بدقة التجارب الكمية [18](Wolberg. 2010). ويعرف أيضا بأنه العملية التي تكون فيها النتيجة على أساس البيانات الحالية. على سبيل المثال، بناء على معلومات الطقس الحالية سيكون تحليل ذلك اليوم يمكن أن يكون إما "مشمسًا" أو "ممطرًا" أو "غائم". [19](Chakarverti. 2019)

يتم اتباع خطوتين في هذه العملية. هم: [20](Chakarverti. 2019)

أ. نموذج البناء: يشرح نموذج البناء مجموعة فئات محددة سلفا. يتم استخدام عدد كبير من المجموعات في ملف بناء النموذج المعروف باسم مجموعة التدريب. يظهر تصنيف القواعد أو أشجار القرار أو الصيغ الرياضية / الانحدار بتنسيق هذه الطريقة.

ب. استخدام النموذج: الطريقة الثانية المستخدمة في ملف التصنيف هو استخدام النموذج. تستخدم نتيجة تصنيف النموذج للمقارنة في اختبار العينة مع الملصق المعروف. مجموعة الاختبار لا تعتمد على عدة التدريبات.

[18] 14- J. Wolberg. Prediction Analysis in Designing Quantitative Experiments. 2010
[19] 12- Mohini Chakarverti, Nikhil Sharma and Rajiva Ranjan Divivedi. Prediction Analysis Techniques of Data Mining: A Review. 2019
[20] 12- Mohini Chakarverti, Nikhil Sharma and Rajiva Ranjan Divivedi. Prediction Analysis Techniques of Data Mining: A Review. 2019

يسمح استخدام تحليل التنبؤ لمصمم التجربة بتقدير الدقة التي يجب الحصول عليها من التجربة قبل الحصول على المعدات والانتهاء من الإعداد التجريبي. من خلال استخدام هذه الطريقة، يتم تطوير صورة توضح كيف يجب أن تعتمد دقة النتائج على المتغيرات التجريبية. (Wolberg, 2010)[21]

[21] 14- J. Wolberg. Prediction Analysis in Designing Quantitative Experiments. 2010

2-**التنقيب في المعطيات الوصفية:** للعثور على الأنماط التي تصف البيانات. يمكن أيضًا تقسيم مهام التنقيب في المعطيات الوصفية إلى أربعة أنواع كالتالي:

- تحليل العنقدة(Clustering)
- تحليل التلخيص (Summarization)
- تحليل قواعد الرابطة (Association Rules)
- تحليل اكتشاف التسلسل (Sequence Discovery)

مثال عن تطبيق التنقيب في المعطيات الوصفية: كشف الاحتيال

2-1 تحليل العنقدة (Clustering Analysis)

تحليل العنقدة أو ببساطة العنقدة هو عملية تجميع مجموعة من كائنات البيانات في مجموعات أو عناقيد متعددة بحيث يكون للكائنات الموجودة داخل الكتلة تشابه كبير، ولكنها تختلف كثيرًا عن الكائنات في مجموعات أخرى. العنقدة كأداة للتنقيب عن البيانات لها جذورها في العديد من مجالات التطبيق مثل علم الأحياء والأمن والبحث على الويب. (Han et al. 2011)[22]

مجموعة الكتل الناتجة عن تحليل الكتلة يمكن أن يشار إليها باسم الكتلة أو العنقود. في هذا السياق قد تولد طرق تجميع مختلفة مجموعات مختلفة على نفس مجموعة البيانات. لا يتم تنفيذ

[22] 2- Jiawe Han, Micheline Kamber, Jian Pei. Data mining: Concepts and Techniques ,3rd ed. (Morgan Kaufmann publisher,2011).

التقسيم يدويّا، ولكن بواسطة خوارزمية العنقدة. ومن ثم، العنقدة مفيد في أنه يمكن أن يؤدي إلى اكتشاف مجموعات غير معروفة سابقًا داخل البيانات. (Han et al. 2011)

استعمالات تحليل العنقدة: [23](Han et al. 2011)

في ذكاء الأعمال، يمكن استخدام العنقدة لتنظيم عدد كبير من العملاء في مجموعات، حيث يشترك العملاء داخل المجموعة في خصائص قوية متشابهة. هذا يسهل تطوير استراتيجيات الأعمال لتحسين إدارة العلاقات مع العملاء.

في التعرف على الصور، يمكن استخدام التجميع لاكتشاف المجموعات أو "الفئات الفرعية" في أنظمة التعرف على الحروف المكتوبة بخط اليد. لنفترض أن لدينا مجموعة بيانات أرقام مكتوبة بخط اليد، حيث يتم تصنيف كل رقم على أنه إما 1، 2، 3، وهكذا. لاحظ أنه يمكن أن يكون هناك اختلاف كبير في الطريقة التي يكتب بها الأشخاص نفس الرقم. خذ الرقم 2 على سبيل المثال. قد يكتبها بعض الأشخاص بدائرة صغيرة في الجزء السفلي الأيسر، بينما يكتب البعض الآخر بدون تلك الدائرة. يمكننا استخدام التجميع لتحديد الفئات الفرعية لـ "2"، كل منها يمثل الاختلاف في الطريقة التي يمكن بها كتابة 2. استخدام نماذج متعددة استنادًا إلى الفئات الفرعية يمكن أن يحسن دقة التعرف الإجمالية.

[23] 2- Jiawe Han, Micheline Kamber, Jian Pei. Data mining: Concepts and Techniques ,3rd ed. (Morgan Kaufmann publisher,2011).

وجدت العنقدة أيضًا في العديد من التطبيقات في بحث الويب. على سبيل المثال، عند البحث باستخدام كلمة واحدة، يعرض البحث الكثير من النتائج (أي الصفحات ذات الصلة بالبحث) نظرًا للعدد الكبير جدًا من صفحات الويب. يمكن استخدام التجميع لتنظيم نتائج البحث في مجموعات وتقديم النتائج بطريقة موجزة ويسهل الوصول إليها. علاوة على ذلك، تم تطوير تقنيات العنقدة لتجميع الوثائق في موضوعات، التي تستخدم عادة في ممارسة استرجاع المعلومات.

2-1-1 تقنيات التجميع

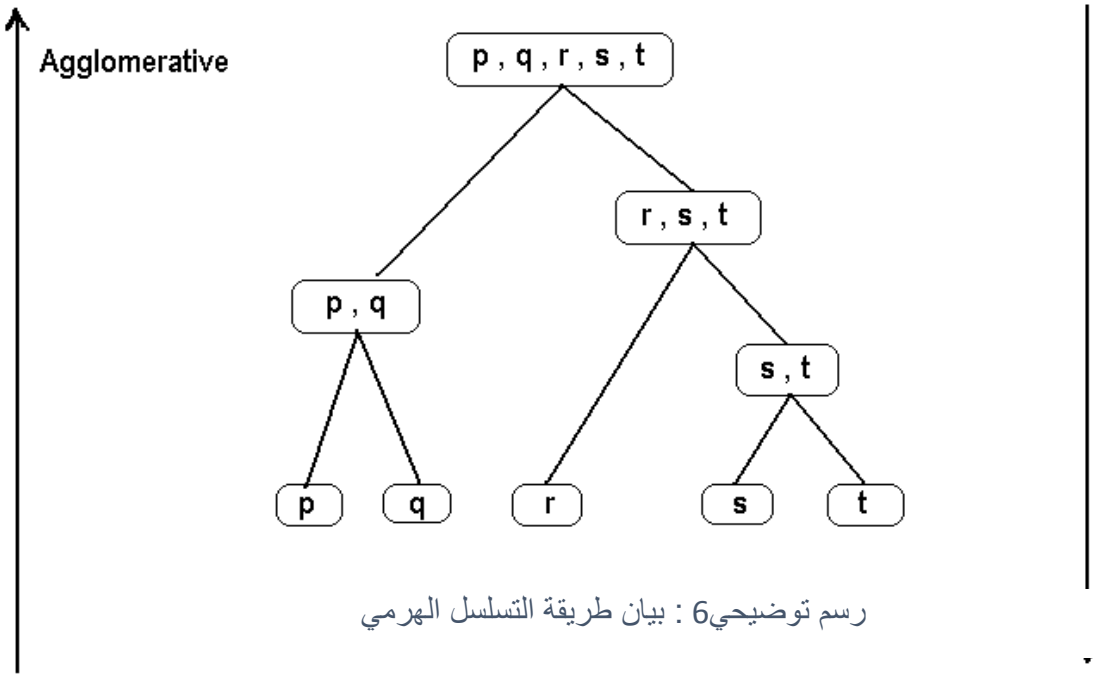
تعتمد معظم خوارزميات التجميع على تقنيتين شائعتين تعرفان بالتسلسل الهرمي (Hierarchical Method) وطريقة التقسيم(Partitioning Method) . (Omran et al. 2007)[24]

ولكن يوجد غيرهم الكثير؛ كالأسلوب القائم على الكثافة (Density-Based Method) والأسلوب القائم على الشبكة (Grid-Based Method)والأسلوب القائم على النموذج Model-Based (Method) والأسلوب القائم على القيد(Constraint-Based Method) . (Han et al. 2011)[25]

[24] 15- Mahamed G.H. Omran, Andries P. Engelbrecht and Ayed Salman. An overview of clustering methods. 2007
[25] 2- Jiawe Han, Micheline Kamber, Jian Pei. Data mining: Concepts and Techniques ,3rd ed. (Morgan Kaufmann publisher,2011).

طريقة التسلسل الهرمي(Hierarchical Method) (Han et al. 2011)

الطريقة الهرمية تخلق تحليلاً هرميًا لمجموعة معينة من كائنات البيانات. يمكن تصنيف الطريقة الهرمية على أنها إما تكتليةاتراكمية(Agglomerative) أو مثيرة للانقسام(Divisive) ، بناءً على كيفية تكوين التحلل الهرمي. يبدأ النهج التراكمي، ويسمى أيضًا النهج التصاعدي، بكل كائن يشكل مجموعة منفصلة. يدمج على التوالي الكائنات أو المجموعات القريبة لبعضها البعض، حتى يتم دمج جميع المجموعات في مجموعة واحدة (المستوى الأعلى من التسلسل الهرمي)، أو شرط الإنهاء معلق. النهج الانقسام، ويسمى أيضا نهج من أعلى إلى أسفل، يبدأ بجميع الكائنات في نفس المجموعة. في كل متتالية التكرار، يتم تقسيم الكتلة إلى مجموعات أصغر، حتى يصبح كل كائن في النهاية في عنقود على حدى، أو عقد شرط الإنهاء.



رسم توضيحي6 : بيان طريقة التسلسل الهرمي

طريقة التقسيم (Partitioning Method): (Han et al. 2011)[26]

بالنظر إلى مجموعة من العناصر n، فإن طريقة التقسيم تنشأ k أقسام(Partitions) البيانات، حيث يمثل كل قسم عنقود و $k \leq n$. هذا هو يقسم البيانات إلى مجموعات k بحيث يجب أن تحتوي كل مجموعة على عنصر واحد على الأقل. بمعنى آخر، تقوم طرق التقسيم بإجراء تقسيم على مستوى واحد على مجموعات البيانات.

[26] 2- Jiawe Han, Micheline Kamber, Jian Pei. Data mining: Concepts and Techniques ,3rd ed. (Morgan Kaufmann publisher,2011).

لدى طريقة التقسيم خوارزميتين أساسيتين هم K-Means وK-Medoids:

تحدد خوارزمية K-Means النقطة الوسطى من العنقود كقيمة متوسطة للنقاط داخل الكتلة.[2]

والخوارزمية تكون كالتالي: [27](Han et al. 2011)

مدخل:

K: عدد المجموعات ،

D : مجموعة بيانات تحتوي على عدد n كائنات.

الإخراج: مجموعة من k عناقيد.

طريقة:

(1) اختيار كائنات k بشكل عشوائي من D كمراكز المجموعة الأولية؛

(2) كرر

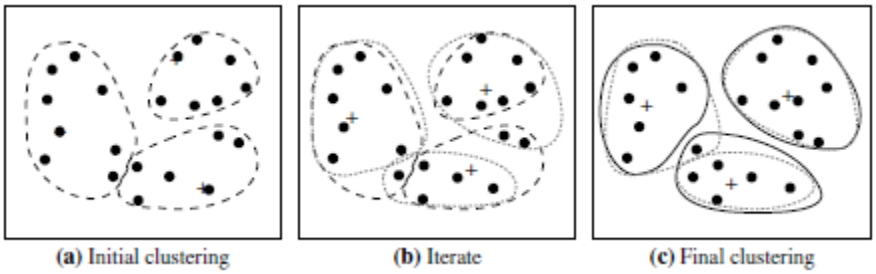
[27] 2- Jiawe Han, Micheline Kamber, Jian Pei. Data mining: Concepts and Techniques ,3rd ed. (Morgan Kaufmann publisher,2011).

(3) (إعادة) تخصيص كل عنصر للعنقود الذي يكون الكائن هو الأكثر تشابهًا معه، على أساس القيمة المتوسطة للعناصر في العنقود؛

(4) تحديث متوسط العنقود، أي حساب متوسط قيمة العناصر لكل كتلة

(5) إعادة حتى لا تغيير؛

تم تلخيص إجراء K-Means في الشكل 7



رسم توضيحي7 : طريقة K-Means

تُستخدم خوارزمية K-Medoids للعثور على Medoids في مجموعة تقع في مركز نقطة في العنقود. K-Medoids أكثر قوة مقارنة ب- K-Means كما هو الحال في K–Medoids، نجد k ككائن تمثيلي لتقليل مجموع الاختلافات في كائنات البيانات. بينما، تستخدم K–Means

مجموع المسافات الإقليدية المربعة للبيانات أشياء. ويقلل مقياس المسافة هذا الضوضاء والقيم المتطرفة. [28](Arora et al. 2015)

والخوارزمية تكون كالتالي: [29](Arora et al. 2015)

الإدخال:

K: عدد الكتل ،

D: مجموعة بيانات تحتوي على n عناصر.

الإخراج: مجموعة من مجموعات k.

الخوارزمية:

- حدد k عشوائيًا باعتباره Medoids لـ n من نقاط البيانات.
- ابحث عن أقرب Medoids عن طريق حساب المسافة بين نقطتي البيانات n و k Medoids وعين عناصر البيانات لذلك.
- لكل Medoids m وكل نقطة بيانات o مرتبطة بـ m قم بما يلي:

[28] 16-2015 Preeti Arora, Dr. Deepali, Shipra Varshney. Analysis of K-Means and K-Medoids Algorithm for Big Data.

[29] 16-2015 Preeti Arora, Dr. Deepali, Shipra Varshney. Analysis of K-Means and K-Medoids Algorithm for Big Data.

1. قم بتبديل m و o لحساب التكلفة الإجمالية للتكوين
2. حدد o Medoids بأقل تكلفة للتكوين.

- إذا لم يكن هناك تغيير في التخصيصات، كرر الخطوتين 2 و3 بدلاً من ذلك

2-2 تحليل التلخيص (Summarization)

يعتبر التلخيص بمثابة منهج أساسي لاكتشاف المعرفة ينتج عنه نسخة موجزة وغنية بالمعلومات من مجموعة البيانات الأصلية.

هناك نوعان رئيسيان من التلخيص؛ الدلالي والنحوي. يتضمن التلخيص النحوي تقنيات البرمجة التقليدية التي تعتبر مجموعة البيانات بأكملها على أنها سلسلة من البايت (Byte) وتستخدم في الغالب لضغط البيانات باستخدام نهج المعلومات النظري. يجد تلخيص البيانات الدلالية وصفًا مضغوطًا لمجموعة البيانات الأصلية مع الحفاظ على بنية مجموعة البيانات الأصلية. (Ahmed et al. 2014)[30]

2-3 تحليل قواعد الرابطة (Association Rules)

نشأ تعدين قواعد الرابطة في تحليل سلة السوق الذي يهدف إلى فهم سلوك واهتمامات التسوق لعملاء التجزئة. يساعد هذا المفهوم في كيفية وضع المنتج (Product Placement) والتسويق

[30] 17- Mohiuddin Ahmed, Abdum Naser Mahmud. Clustering Based Semantic Data Summarization Technique: A New Approach. 2014

المباشر(Direct Marketing) . يكتشف تحليل قواعد الرابطة الأنماط في سلال السوق المرصودة والتي تحدث بشكل متكرر. [31](Hegland. M. 2003)

يمكن العثور على نوعين من الأنماط في تعدين قواعد الارتباط: النوع الأول هو:

"if–then–rules" وهي بالشكل: "إذا اشترى الزبون الحليب ، فإنه يشتري الخبز أيضًا".

النوع الثاني يتعلق بالتواجد المشترك للعناصر في سلة السوق: "يشتري العميل الخبز والحليب معًا". اكتشاف النمط الثاني أبسط من الأول، علاوة على ذلك، يمكن للمرء أن يرى أن اكتشاف النمط الأول يمكن أن يعتمد على اكتشاف النمط الثاني. [32](Hengland M. 2003)

أحد أهم الخوارزميات المتبعة والتي ستعمل عليها الباحثة في هذا المشروع هي خوارزمية ابريوري (Apriori). كما أن هذه الخوارزمية من أهم الخوارزميات البسيطة المتبعة في تعدين مجموعات العناصر المتكررة في مجموعة من المعاملات.

[31] 18- Hegland, M. Algorithms for Association Rules. 2003
[32] 18- Hegland, M. Algorithms for Association Rules. 2003

3- عملية اكتشاف المعرفة في قواعد البيانات (KDP)

وهي عملية العثور على المعرفة في البيانات، ويتم ذلك باستخدام طرق التنقيب في المعطيات (الخوارزميات) من أجل استخراج المعرفة المطلوبة من كمية كبيرة من البيانات.

وتتكون العملية من الخطوات التالية:

1-تنظيف البيانات (Data Cleaning): لإزالة الضجيج والبيانات غير متناسقة.

2-تكامل البيانات (Data Integration): حيث يمكن الجمع بين مصادر بيانات متعددة.

3-اختيار البيانات (Data Selection): حيث يتم اختيار البيانات ذات صلة لهذه المهمة من قاعدة البيانات.

4-تحول البيانات (Data transformation): حيث يتم تحويل البيانات الى أشكال مناسبة.

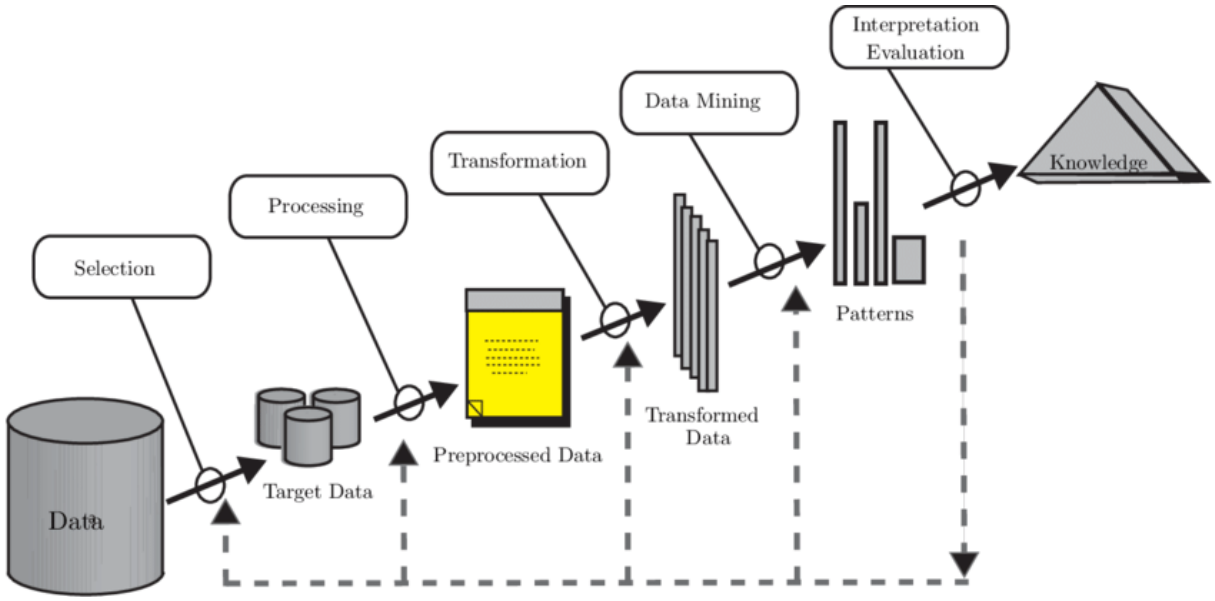
5-التنقيب في البيانات (Data Mining): عملية أساسية لتطبيق طرق ذكية لاستخراج أنماط البيانات.

6-تقييم الأنماط (Pattern Evaluation): حديد التعرف على أنماط مثيرة للاهتمام لتمثيل المعرفة.

7-تمثيل المعرفة (Knowledge Presentation): تقنيات لتمثيل وتصوير المعرفة، وتستخدم لتقديم المعرفة المفيدة للمستخدم.

الخطوات من 1 إلى 4 تمثل أشكال مختلفة لتجهيز البيانات، حيث يتم إعدادها للتنقيب. فتشمل دمج البيانات الواردة من عدة مصادر وتنقية البيانات لإزالة التشويش (Noise).

وباعتبار أن هنالك طرق عديدة لجمع وتخزين البيانات، فإن المعالجة المسبقة للبيانات قد تكون أكثر الخطوات إجهاداً واستهلاكاً للوقت من بين خطوات عملية اكتشاف المعرفة. استخراج البيانات هو عملية اكتشاف أنماط مثيرة للاهتمام ومعرفة من كميات كبيرة من البيانات. ويمكن أن تشمل مصادر البيانات قواعد البيانات ومستودعات البيانات الويب ومصادر اخرى. (Han et al. 2011)[33]



رسم توضيحي8 : عملية اكتشاف المعرفة

[33] 2- Jiawe Han, Micheline Kamber, Jian Pei. Data mining: Concepts and Techniques ,3rd ed. (Morgan Kaufmann publisher,2011).

4- ما نوع البيانات التي ممكن استخدامها في تنقيب في المعطيات؟

كتقنية عامة، يمكن تطبيق التنقيب في المعطيات على أي نوع من البيانات طالما أن البيانات ذات مغزى للتطبيق الهدف. أبسط أشكال البيانات للتعدين التطبيقات هي بيانات قاعدة البيانات، وبيانات مستودع البيانات، وبيانات المعاملات. [34](Han et al. 2011)

بيانات قاعدة البيانات:

يتكون نظام قاعدة البيانات، المعروف أيضًا باسم نظام إدارة قواعد البيانات (DBMS)، من مجموعة من البيانات المترابطة، والمعروفة باسم قاعدة البيانات، ومجموعة من البرامج لإدارة والوصول إلى البيانات. توفر البرامج آليات للتعريف هياكل قواعد البيانات وتخزين البيانات؛ لتحديد وإدارة المتزامنة والمشتركة أو الوصول إلى البيانات الموزعة؛ ولضمان اتساق وأمن المعلومات المخزنة على الرغم من أعطال النظام أو محاولات الوصول غير المصرح به. (Han et al. 2011)

قاعدة البيانات العلاقاتية هي عبارة عن مجموعة من الجداول، يتم تخصيص جدول فريد لكل منها اسم. يتكون كل جدول من مجموعة من السمات (أعمدة أو حقول) وعادة ما يتم تخزينها مجموعة كبيرة من المجموعات (سجلات أو صفوف). كل مجموعة في جدول علائقي تمثل ملف كائن مُعرَّف بمفتاح فريد ويتم وصفه بواسطة مجموعة من قيم السمات. دلالي غالبًا ما يتم إنشاء

[34] 2- Jiawe Han, Micheline Kamber, Jian Pei. Data mining: Concepts and Techniques ,3rd ed. (Morgan Kaufmann publisher,2011).

نموذج البيانات، مثل نموذج بيانات علاقة كيان (ER) قواعد البيانات العلائقية. يمثل نموذج بيانات التقارير الإلكترونية قاعدة البيانات كمجموعة من الكيانات وعلاقاتهم.

يمكن الوصول إلى البيانات العلائقية عن طريق استعلامات قاعدة البيانات المكتوبة بلغة استعلام علائقي (على سبيل المثال، SQL) أو بمساعدة واجهات المستخدم البيانية. عند تعدين قواعد البيانات العلائقية، يمكننا المضي قدمًا بالبحث عن الاتجاهات أو أنماط البيانات. على سبيل المثال، يمكن لأنظمة التنقيب عن البيانات تحليل بيانات العملاء للتنبؤ مخاطر الائتمان للعملاء الجدد بناءً على دخلهم وعمرهم وائتمانهم السابق معلومة. [35](Han et al. 2011)

مستودع البيانات:

مستودع بيانات هو مستودع للمعلومات التي تم جمعها من مصادر متعددة، ومخزنة في إطار مخطط موحد، ويقيمون عادةً في موقع واحد. يتم إنشاء مستودعات البيانات عبر عملية تنظيف البيانات وتكامل البيانات وتحويل البيانات وتحميل البيانات وتحديث الدوري للبيانات. (Han et al. 2011)

لتسهيل اتخاذ القرار، يتم تنظيم البيانات الموجودة في مستودع البيانات حول الموضوعات الرئيسية (على سبيل المثال، العميل، والعنصر، والمورد، والنشاط). يتم تخزين البيانات لتقديم معلومات

[35] 2- Jiawe Han, Micheline Kamber, Jian Pei. Data mining: Concepts and Techniques ,3rd ed. (Morgan Kaufmann publisher,2011).

من منظور تاريخي، كما في الأشهر الستة إلى الاثني عشر الماضية، ويتم بعدها تلخيصهم عادة. على سبيل المثال، بدلاً من تخزين تفاصيل كل معاملة مبيعات، قد يخزن مستودع البيانات ملخصًا للمعاملات لكل نوع صنف لكل نوع متجر أو، تلخيصًا إلى مستوى أعلى، لكل منطقة مبيعات.

بيانات المعاملات:

يلتقط كل سجل في قاعدة بيانات المعاملات معاملة، مثل شراء العميل أو حجز رحلة طيران أو نقر المستخدم على صفحة ويب. تتضمن المعاملة غالبا رقم هوية فريدة (trans_ID) وقائمة بالعناصر الواصفة للمعاملة، مثل العناصر المشتراة في المعاملة. قد تحتوي قاعدة البيانات المعاملة على جداول إضافية، والتي تحتوي على معلومات أخرى ذات صلة بالمعاملات، مثل وصف العنصر أو معلومات حول مندوب المبيعات أو فرع، وهكذا. [36](Han et al. 2011)

ويكون شكل جدول المعاملات كالتالي:

[36] 2- Jiawe Han, Micheline Kamber, Jian Pei. Data mining: Concepts and Techniques ,3rd ed. (Morgan Kaufmann publisher,2011).

trans ID	list of item IDs
T100	I1, I3, I8, I16
T200	I2, I8
T300	I1, I3, I5
….	…..

أنواع أخرى من البيانات:

إلى جانب بيانات قاعدة البيانات العلائقية وبيانات مستودع البيانات وبيانات المعاملات، هناك العديد من أنواع البيانات الأخرى التي لها أشكال وهياكل متعددة الاستخدامات ومختلفة نوعًا ما المعاني الدلالية. يمكن رؤية مثل هذه الأنواع من البيانات في العديد من التطبيقات: مرتبطة بالوقت أو بيانات التسلسلية (السجلات التاريخية، وبيانات البورصة، والسلاسل الزمنية)، وتدفقات البيانات (بيانات المراقبة بالفيديو وأجهزة الاستشعار، وهي تنتقل باستمرار)، البيانات المكانية (مثل الخرائط)، والرسم البياني والبيانات المتصلة بالشبكة (الشبكات الاجتماعية والمعلوماتية)، والويب. تجلب هذه التطبيقات تحديات جديدة، مثل كيفية التعامل مع البيانات التي تحمل هياكل

خاصة (مثل التسلسلات والأشجار، والرسوم البيانية والشبكات) وكيفية تعدين تلك الأنماط. (Han et al. 2011)[37]

5- ما هي قضايا التنقيب في المعطيات؟ (R. Mohammed. 2014)[38]

- **القضايا التجارية:** هي عملية تحليل البيانات التجارية الموجهة وتحويلها وتصنيفها.
- القضايا الاجتماعية.
- **قضايا منهجية التنقيب:** مناسبة وملائمة لتنقيب البيانات المطبقة وقيودها.
- **التكلفة:** بينما انخفضت تكاليف أجهزة النظام بشكل كبير خلال السنوات القليلة الماضية، يميل التنقيب عن البيانات وتخزين البيانات إلى تعزيز الذات.
- **قضايا واجهات المستخدم:** المعرفة المكتشفة بواسطة أدوات التنقيب عن البيانات مفيدة طالما أنها مثيرة للاهتمام وقبل كل شيء مفهومة من قبل المستخدم.
- **قضايا مصادر البيانات:** تظهر فائض من البيانات عندما يكون لدينا بيانات أكثر مما نتحمله – يتم تخزين أنواع مختلفة من البيانات في مجموعة متنوعة من المستودعات.

[37] 2- Jiawe Han, Micheline Kamber, Jian Pei. Data mining: Concepts and Techniques ,3rd ed. (Morgan Kaufmann publisher,2011).
[38] 5- Radwan Mohammed. Data Mining, 2014

القسم الثاني

خوارزمية Apriori ونظم التوصية

1- خوارزمية أبريوري (Apriori Algorithm)

تعد خوارزمية Apriori واحدة من أكثر الطرق شهرة للتنقيب عن مجموعات العناصر المتكررة (Frequent Itemsets) في قاعدة بيانات المعاملات. تعمل الخوارزمية ضمن إطار إنشاء واختبار متعدد التمريرات، يشتمل على مرحلتي الانضمام والتقليم إلى تقليل عدد المرشحين قبل مسح قاعدة البيانات لحساب الsupport . [39](M. Yen-Lin et al. 2012)

تقوم خوارزمية Apriori بتجميع جميع العناصر المتكررة في قاعدة بيانات المعاملات (Transactions)، حيث تحتوي كل معاملة على مجموعة من العناصر تسمى Itemset. تكون مجموعة العناصر X متكررة إذا كان دعمها على الأقل حدًا أدنى معينًا للدعم يحدده المستخدم min_sup. (M. Yen-Lin et al. 2012) حد أدنى للدعم أو Minimum (min_sup) Support هو حد أدنى لتكرارات عنصر ما بين مجموعة معاملات. تسمى مجموعة العناصر التي تحتوي على عناصر k (ذو العناصر بدعم أعلى من الmin_sup) مجموعة عناصر k وطولها k.

اذاً تبدأ الخوارزمية بتحديد ال min_sup (قد يكون 1 او أكثر بحسب حاجة الباحث). ثم يتكون مفهوم خوارزمية apriori من عدة مراحل تكرار، كل تكرار سيولد نمط بيانات مع أساس البيانات للحصول على الدعم من كل عنصر. سيتم تحديد العناصر التي لها دعم أعلى من الحد الأدنى

[39] 19- Ming-Yen Lin, Pei-Yu Lee, Sue-Chen Hsueh. Apriori-based Frequent Itemset Mining Algorithms on MapReduce. 2012

للدعم وتصبح نمطًا عالي التردد يسمى itemset 1. في عملية التكرار الثانية، ستنتج عنصرين يتم تحديدهما من قبل كل طرف لاحتوائهما على عنصرين. تستمر العملية حتى لا يوجد حد أدنى من الدعم. (Chandra et al. 2019)[40]

الدعم Support هو مقياس كبير يهيمن على العناصر من الصفقة بأكملها. الثقة Confidence هي مقياس يوضح العلاقة بين عنصرين بشكل مشروط بناءً على شروط معينة. (Chandra et al.2019[41] أي عدد المرات التي يتم فيها شراء المنتجين معًا مثلا. الرفع Lift هو ما يعطينا التكرار المستقل لـ X و Y. (A. Das. 2021)[42] يوضح هذا مدى احتمالية شراء العنصر Y عند شراء العنصر X، مع التحكم في مدى شيوع العنصر Y.

$$Support = \frac{frq(X,Y)}{N}$$

$$Rule: X \Rightarrow Y \longrightarrow Confidence = \frac{frq(X,Y)}{frq(X)}$$

$$Lift = \frac{Support}{Supp(X) \times Supp(Y)}$$

رسم توضيحي9 : قواعد الدعم والثقة والرفع

[40] 20- J Chandra, K R Dewi. Implementation of Data Mining Sales of Milk Using Apriori Algorithm Method. 2019
[41] 20- J Chandra, K R Dewi. Implementation of Data Mining Sales of Milk Using Apriori Algorithm Method. 2019
[42] 21- Arijita Das, Soumita Jana, Pranita Gangouly. Application of Association Rule:Apriori Algorithm in E-Commerce. 2021

الشكل رقم 9 يوضح قوانين الدعم والثقة والرفع حيث Xو Y عنصرين في مجموعة معاملات ما يتم العمل عليهم لتحصيل النتائج.

قدم Agrawal & Srikant طريقة Apriori في عام 1994 لتحديد مجموعات العناصر المتكررة لقواعد الارتباط المنطقية. في التنقيب عن البيانات، تعد خوارزمية apriori نوعًا من قواعد الارتباط. تحليل التقارب(Affinity Analysis) أو تحليل سلة السوق (Market Basket Analysis) هي المصطلحات المستخدمة لوصف القواعد التي تعبر عن العلاقة بين العديد من الصفات. يتم استخدام نهج التنقيب عن البيانات لتحليل الارتباط، والمعروف أيضًا باسم التنقيب عن قواعد الارتباط(Association Rule Mining) ، لاكتشاف قواعد مجموعة من العناصر. (P. Edastama et al. 2021)[43]

2-1 طرق وتقنيات تحسين لخوارزمية Apriori

- **طريقة تقسيم البيانات Data Partitioning Method:** من أجل تقليل عدد عمليات مسح قاعدة البيانات، اقترح Savasere طريقة التقسيم على أساس التقسيم. تنقسم قاعدة البيانات إلى العديد من الأقسام، والفكرة الأساسية هي: مجموعات العناصر المتكررة في قاعدة البيانات على الأقل جزء واحد في قاعدة البيانات متكرر،

[43] 23- Primasatria Edastama, Ankur Singh Bist, Ari Prambudi. Implementation of Data Mining on Glasses Sales Using the Apriori Algorithm. 2021

ثم مجموعة الاتحاد من مجموعات العناصر المتكررة لكل مقطع هي مجموعة العناصر المتكررة المحتملة. [44](Z. Chen. 2011)

- **تقنية مبنية على التجزئة Hash–Based Technique:** يستخدم بنية بيانات التي تمثل جدول تجزئة مباشرة. تقترح هذه الخوارزمية التغلب على بعض نقاط الضعف في خوارزمية Apriori عن طريق تقليل عدد مجموعات العناصر المرشح. على وجه الخصوص، المجموعة الثانية من العناصر، لأن هذا هو مفتاح تحسين الأداء. تستخدم هذه الخوارزمية تقنية تعتمد على التجزئة لتقليل عدد مجموعات العناصر المرشحة التي تم إنشاؤها في المرور الأول. يُزعم أن عدد مجموعات العناصر في C2 التي تم إنشاؤها باستخدام التجزئة يمكن تصغيرها، بحيث يكون الفحص المطلوب لتحديد L2 أكثر كفاءة. [45](S. Shakya. 2013)
- **اختبار العينات Sampling:** الفكرة الأساسية لهذه الطريقة هي اختيار عينة عشوائية S من البيانات المعطاة D، ثم البحث عن مجموعات العناصر المتكررة في S بدلاً من D. لأننا نبحث عن مجموعات متكررة في S بدلاً من D، فمن الممكن أننا سنفتقد

[44] 24- Zhuang Chen, Shibang Cai, Qiulin Song and Chonglai Zhu. An Improved Apriori Algorithm Based on Pruning Optimization and Transaction Reduction. 2011

[45] 25- Santosh Shakya, Anju Singh, Divakar Singh. A Survey on Hash based A-priori Algorithm for Web Log Analysis. 2013

بعض العناصر المتكررة. لتقليل هذا الاحتمال، نستخدم حد دعم أقل من الحد الأدنى للدعم للعثور على مجموعات العناصر المتكررة المحلية لـ S. [46](Z. Chen. 2011)

[46] 24- Zhuang Chen, Shibang Cai, Qiulin Song and Chonglai Zhu. An Improved Apriori Algorithm Based on Pruning Optimization and Transaction Reduction. 2011

2- أنظمة التوصية Recommendation Systems

تعرّف نظم أو محركات التوصية على أنها أدوات وتقنيات برمجية تساعد وتعزز عملية صنع القرار من خلال تجميع آراء الناس وتوجيههم إلى المعطيات أو العناصر التي تكون غالبا مهمة ومفيدة لهؤلاء الناس. أنظمة التوصية موجهة بشكل أساسي إلى الأفراد الذين يفتقرون إلى الخبرة الشخصية الكافية أو الكفاءة من أجل تقييم العدد الهائل المحتمل للعناصر البديلة التي قد يقدمها موقع ويب.

في أي حال، كتصنيف عام، تشير البيانات التي تستخدمها أنظمة التوصية إلى ثلاثة أنواع من الكائنات: العناصر والمستخدمون والمعاملات، أي العلاقات بين المستخدمين والعناصر. (F. Ricci. 2005)[47]

العناصر: العناصر هي الكائنات الموصى بها. قد تتميز العناصر بتعقيدها وقيمتها أو فائدتها. قد تكون قيمة العنصر إيجابية إذا كان العنصر مفيدًا للمستخدم، أو سلبية إذا كان العنصر غير مناسب واتخذ المستخدم قرارًا خاطئًا عند تحديده. يمكن تمثيل العناصر باستخدام طرق مختلفة للمعلومات والتمثيل. على سبيل المثال، بطريقة مبسطة كرمز معرف واحد، أو في شكل أكثر ثراءً، كمجموعة من السمات.

[47] 26- Francesco Ricci, Lior Rokach, Bracha Shapira. Recommender Systems Handbook. 2015

المستخدمين: قد يكون لمستخدمي نظام التوصية أهداف وخصائص متنوعة للغاية. من أجل إضفاء الطابع الشخصي على التوصيات والتفاعل بين الإنسان والحاسوب، تستغل أنظمة التوصية مجموعة من المعلومات حول المستخدمين التي تجمعها من مشاهداتهم وتقييماتهم.

المعاملات: نشير بشكل عام إلى المعاملة على أنها تفاعل مسجل بين المستخدم ونظام التوصية. المعاملات عبارة عن بيانات شبيهة بالسجلات تخزن المعلومات المهمة التي يتم إنشاؤها أثناء التفاعل بين الإنسان والحاسوب والتي تكون مفيدة لخوارزمية إنشاء التوصيات التي يستخدمها النظام.

تقسم نظم التوصية عادة الى عدة تقنيات:(N. Nassar. 2019) [48]

- نظم التوصية المعتمدة على الترشيح التعاوني Collaborative Filtering
- نظم التوصية المعتمدة على المحتوى Content-Based
- نظم التوصية المعتمدة على الثقة Trust-Based
- نظم التوصية المعتمدة على المعرفة Knowledge-Based

نظام التوصية القائم على المحتوى: يتعلم النظام التوصية بالعناصر المشابهة لتلك التي أحبها المستخدم في الماضي. يتم حساب تشابه العناصر بناءً على الميزات المرتبطة بالعناصر المقارنة.

[48] 27- Nour Nassar. Recommender System in Big Data. 2019

على سبيل المثال، إذا قام المستخدم بتقييم إيجابي لفيلم ينتمي إلى النوع الكوميدي، فيمكن للنظام أن يتعلم التوصية بأفلام أخرى من هذا النوع.[49](F. Ricci. 2015)

نظام التوصية المعتمد على الترشيح التعاوني: يقدم التطبيق الأصلي والأكثر بساطة لهذا الأسلوب توصيات للمستخدم النشط بناءً على العناصر التي أحبها المستخدمون الآخرون ذوو الأذواق المماثلة في الماضي. يعتبر الترشيح التعاوني هو الأسلوب الأكثر شيوعًا والمنفذ على نطاق واسع في نظام التوصيات. (F. Ricci. 2015)

نظام التوصية المعتمد على المعرفة: توصي الأنظمة القائمة على المعرفة بالعناصر بناءً على معرفة المجال المحددة حول كيفية تلبية ميزات عنصر معينة لاحتياجات المستخدمين وتفضيلاتهم، وفي النهاية، كيف يكون العنصر مفيدًا للمستخدم. (F. Ricci. 2015)

نظام التوصية المعتمد على الثقة: في نظام التوصية، تم وصف الثقة على أنها: "الاعتقاد المحلي لمستخدم واحد في فائدة التوصية المقدمة من قبل مستخدم آخر". كما تم تعريف أنظمة التوصية القائمة على الثقة على أنها أنظمة تعاونية تعتمد على علاقات المستخدم التي تعبر عن الثقة فيما بينها.[50](S.Afif et al. 2016)

[49] 26- Francesco Ricci, Lior Rokach, Bracha Shapira. Recommender Systems Handbook. 2015
[50] 28- Selma Afif, Zaki Brahmi, Mohamed Mohsen Gammoudi. Trust-Based Recommender System: An Overview. 2016

الفصل الثالث

نظام CMoRS

نظام كوكي مونستر للتوصية

لمحة عامة عن محل مبيع الحلويات Cookie Monster

كوكي مونستر هو مشروع عمل بيتي صغير افتتحته باحثة هذا المشروع بكانون الأول من عام 2020 لبيع الحلويات وبالأخص بسكويت الأميركي المسمى بالكوكيز.

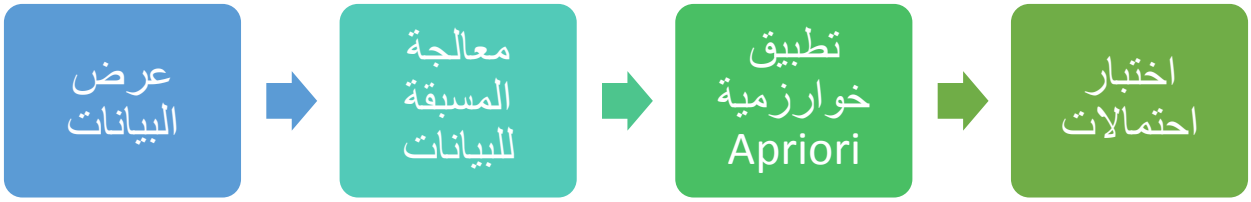
بدأت ب 4 أنواع من البسكويت فقط، وبت أضيف أنواع ونكهات جديدة منذ ذلك الحين. اليوم لدي ما يقارب ال 20 نوع من الكوكيز وزبائن من مختلف محافظات سوريا. كما أني أبيع عن طريق شركة BeeOrder التي تختص ببيع وتوصيل الطعام في دمشق.

تذكير بالسبب

إن أهم والهدف الأساسي من الشركات والأعمال هو الربح، وقد ثبت مرارا وتكرارا أن التركيز على العميل هو العامل الأساسي لهذا الربح. لذلك، فإن الهدف من البحث تصميم نظام توصية يساعد الزبائن والمستخدمين لاختيار أنسب المواد التي تلائم رغباتهم، مما قد يساعد من رفع أرباح المحل.

نموذج الحل (طريقة الحل)

يبدأ الحل بعرض البيانات المجمعة، وهي المعاملاتTransaction، عن طريق رماز مصدري مكتوب بلغة Python، بعدة طرق. في المرحلة الأولى، عرضت المعاملات كما هي في الجدول، ثم تم تحويلها الى مصفوفات من 195 سطر ليتم بعد ذلك جمع أنواع المواد الموجودة ضمن المعاملات في جدول يتضمن اسم النوع وعدد تكراره ضمن المعاملات. في المرحلة الثانية، وهي المرحلة التي بدأ فيها بناء النظام المقترح، بدأت مرحلة المعالجة المسبقة للبيانات Preprocessing وتحويل صيغة البيانات Data Transformation. حولت البيانات مرة أخرى الى مصفوفات وثم الى جدول قيم المنطقية Boolean Chart. بعد ذلك تم الدخول الى المرحلة التي تليها، وهي التنقيب في المعطيات Data Mining حيث تم استخدام خوارزمية Apriori للحصول على قواعد الارتباط. بعد ذلك، تم تجريب بعض احتمالات استخدام هذا



النظام، من خلال افتراض رغبات الزبون في الشراء ومن ثم تطبيقها على النظام ليوصي لهم أنسب منتج بحسب الرغبة. وكان مجموع الاحتمالات 4، اختبر فيها عدة نماذج من طلبات المحتملة للزبائن

البيانات المجمعة (Collected Data)

إن البيانات المجمعة في هذا البحث والتي ستستخدمها الباحثة هي المعاملات(Transactions) التي أجريت عن طريق شركة Beeorder. بسبب ضياع بعض المعاملات التي حصلت عن طريق البيع من Beeorder، اضطرت لاستخدام البيانات بدءا من الشهر العاشر من عام 2021 الى أواخر الشهر الخامس من عام 2022، أي 8 أشهر، والتي قدرت ب 195 معاملة. تتضمن كل معاملة اسم الشخص وأنواع الكوكيز التي طلبت. ستعمل الباحثة على الكوكيز التي طلبت من كل معاملة وتطبق خوارزمية Apriori عليها، لتحصيل علاقات ترابط بين أنواع الكوكيز التي طلبت وبناء نظام توصية يعتمد على الخوارزمية.

سُجلت المعاملات على ملف اكسل ثم حُوّلت إلى ملف CSV (Comma Separated Values)وهو ملف يحول كل سطر من ملف اكسل الى سطر خطي بفواصل بين القيم. وذلك لتسهيل قراءة الملف من قبل عدة برامج أو في حالتنا، من قبل لغة برمجة Python.

لماذا بايثون Python؟

بايثون هي لغة برمجية أخذت صداها في عقدنا الحالي بسبب بساطة وسهولة تعلم اللغة واستخدامها. تمتاز لغة بايثون بالمكتبات (Libraries) الموجودة فيه التي تسهل على المبرمج أن يكتب الرماز المصدري الخاص لتفعيل خوارزمية ما أو المساعدة في حل معادلات رياضية أو تمثيل بيانات على شكل رسوم بيانية وجداول وغيرها. كما أنها تعتبر الأمثل للاستخدام في مجال التعلم الآلي والذكاء الصناعي بسبب مرونة العمل فيها والتنوع الواسع للمكبات في هذا المجال. وبما أن خوارزمية Apriori تعتبر إحدى خوارزميات التعلم الآلي فإن تطبيقها باستخدام هذه اللغة أفضل. كما أنّها سهلة القراءة بالنسبة لأولئك الذين ليسوا على دراية كافية بها أو للمبتدئين.

الحل المقترح

نبدأ أولا بعرض المكتبات المستخدمة والتي عرضت بالرماز المصدري التالي

```
In [1]: import pandas as pd
        import numpy as np
        from mlxtend.preprocessing import TransactionEncoder
        from mlxtend.frequent_patterns import apriori,association_rules
        import matplotlib.pyplot as plt
```

حيث أن:

Numpy وهي اختصار لNumerical Python تتعامل مع الArrays وتوفر العديد من الشفرات الخاصة بالعمليات الرياضية والمصفوفات وإلخ.

Pandas مكتبة تتعامل مع الكثير من أنواع البيانات، من معالجتها، وتنظيفها، وتحليلها.

Matplotlib مكتبة لتمثيل البيانات (Data Visualization) وتحول البيانات إلى شتى أنواع من الأشكال البيانية والجداول.

Mlxtend مكتبة مختصة بوظائف تحليل البيانات اليومية والتعلم الآلي.

بعد ذلك نعرض أول 10 معاملات لدينا بالرماز المصدري التالي:

```
data = pd.read_csv(r"C:\Users\Dareen\Desktop\Datanew411.csv", header=None)
data.head(10)
```

فنحصل عليها بالشكل التالي:

	0	1	2	3	4	5	6	7	8	9
0	half popcorn	sweet chocolate	brownie	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	offer	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	dark chocolate	peanut butter	smores	walnut	NaN	NaN	NaN	NaN	NaN	NaN
3	full popcorn	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	sweet chocolate	brownie	smores	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5	half popcorn	offer	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
6	peanut butter	smores	offer	NaN	NaN	NaN	NaN	NaN	NaN	NaN
7	half popcorn	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
8	offer	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
9	sweet chocolate	brownie	cinnamon	smores	NaN	NaN	NaN	NaN	NaN	NaN

كما نعلم بأن عدد المعاملات لدينا هي 195 معاملة بالمجموع، للتأكد من الممكن أن نكتب الترميز التالي الذي سيكشف عن عدد المعاملات وعدد العواميد (المواد المباعة في أكبر معاملة).

أي 195 معاملة و 10 عواميد

```
data.shape
```

```
(195, 10)
```

نريد الآن أن نتعرف أكثر على البيانات التي لدينا، على سبيل المثال، ما هي أكثر مادة تم شراءها من بين المعاملات؟ أو ما أقل مادة تم شراءها؟ لعمل ذلك يجب علينا أولا ترتيب المعاملات على شكل مجموعة مصفوفات Arrays. هنا ننشأ متغير سنسميه Transactionفيه سيتم جمع المعاملات على شكل مصفوفات من 195 سطر. ثم نحاولها إلى إطار بيانات Data Frame لإضافة كل العناصر الموجودة في البيانات السابقة إلى المصفوفات حيث أن كل

```
#Gather all items of each transaction into numpy array
transaction =[]
for i in range(0, data.shape[0]):
    for j in range(0, data.shape[1]):
        transaction.append(data.values[i,j])
transaction = np.array(transaction)
#transform to dataframe
df = pd.DataFrame(transaction, columns=["items"])

df["incident_count"] = 1
#delete NaN Values
indexNames = df[df["items"] == "nan" ].index
df.drop(indexNames, inplace=True)
#make new dataframe
df_table = df.groupby("items").sum().sort_values("incident_count", ascending=False).reset_index()
```

مصفوفة تحتوي على 10 أماكن للعناصر وباقي العناصر تبقى Nan. لنتأكد أن قيم الNan لا تدخل بالحل، يجب علينا حذف كل عناصر الNan الموجودة بالرماز Drop.

بعد ذلك نجمع كل العناصر المتشابهة ببعضها ونضعها في جدول يحتوي على كل العناصر. لنرى الآن ما هي اول 10 أكثر العناصر شراء بين كل العناصر. لعمل ذلك، نكتب الكود التالي:

```
df_table.head(10).style.background_gradient(cmap="Blues")
```

والنتيجة تكون كالتالي:

	items	incident_count
0	brownie	86
1	sweet chocolate	73
2	nutella	65
3	half popcorn	52
4	lotus	35
5	milka	34
6	cinnamon	32
7	smores	32
8	full popcorn	24
9	peanut butter	22

حيث أن عنصر Brownie هو الأكثر شراء، فقد تم شراءه 86 مرة.

لنرى الآن ما هي أقل 10 عناصر مبيعا. لعمل ذلك، نكتب الرماز المصدري التالي:

```
df_table.tail(10)
```

فتظهر لدينا النتيجة كالتالي:

	items	incident_count
13	walnut	15
14	dried fruit	14
15	kinder	14
16	m&ms	13
17	offer	13
18	monster	8
19	galaxy	8
20	skittles	4
21	coconut	4
22	maltesers	2

كما نرى أن أقل نوع مبيعا لدينا هو maltesers

لنبدأ الآن بتجهيز البيانات التي لدينا لنطبق عليها خوارزمية Apriori.

أولا يجب أن نحول البيانات الى مصفوفات. ولعمل ذلك نكتب الرماز المصدري التالي:

```
transaction=[]
for i in range(data.shape[0]):
    transaction.append([str(data.values[i,j]) for j in range(data.shape[1])])

transaction = np.array(transaction)
transaction
```

ويصبح لدينا الشكل التالي من البيانات:

```
array([['half popcorn', 'sweet chocolate', 'brownie', ..., 'nan', 'nan',
        'nan'],
       ['offer', 'nan', 'nan', ..., 'nan', 'nan', 'nan'],
       ['dark chocolate', 'peanut butter', 'smores', ..., 'nan', 'nan',
        'nan'],
       ...,
       ['monster', 'sweet chocolate', 'nan', ..., 'nan', 'nan', 'nan'],
       ['kinder', 'galaxy', 'nutella', ..., 'nan', 'nan', 'nan'],
       ['sweet chocolate', 'half popcorn', 'brownie', ..., 'nan', 'nan',
        'nan']], dtype='<U15')
```

ثم نحول هذه المصفوفات الى مخطط منطقي Boolean Chart، بقيم من True أو False.
كما نحذف العمود الذي فيه قيم Nan لتفادي الأغلاط في النتائج النهائية. ولعمل ذلك نكتب الرماز المصدري التالي:

```
te = TransactionEncoder()
te_array= te.fit(transaction).transform(transaction)
dataset = pd.DataFrame(te_array, columns=te.columns_)
dataset.drop('nan', inplace = True, axis=1)
dataset
```

فتكون النتيجة كالتالي:

74

	brownie	cinnamon	coconut	coconut raisin	cream	dark chocolate	dried fruit	full popcorn	galaxy	half popcorn	...	maltesers	milka	monster	nutella	offer	peanut butter	skittles
0	True	False	False	False	False	False	False	False	False	True	...	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	True	False	False
2	False	False	False	False	False	True	False	False	False	False	...	False	False	False	False	False	True	False
3	False	False	False	False	False	False	False	True	False	False	...	False	False	False	False	False	False	False
4	True	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	False	False
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
190	True	False	False	False	False	False	False	False	False	False	...	False	False	False	True	False	False	False
191	False	False	False	False	False	False	False	False	True	False	...	False	False	False	True	False	False	False
192	False	False	False	False	False	False	False	False	False	False	...	False	False	True	False	False	False	False
193	False	False	False	False	False	False	False	False	True	False	...	False	False	True	True	False	False	False
194	True	False	False	False	False	False	False	False	False	True	...	False	False	False	False	False	False	False

195 rows × 23 columns

نبدأ الآن بتطبيق الخوارزمية عبر استدعاءها من مكتبة mlxtend حيث تم "استيراد" import خوارزمية Aprioriوالقواعد الرابطة. نكتب الرماز المصدري التالي:

```
frequent_itemsets = apriori(dataset, min_support=0.07, use_colnames=True)
frequent_itemsets
```

تم اختيار الmin_support بافتراض أنه لدينا طلبيتين في اليوم لمدة 7 أيام. فبضرب 2*7 ثم تقسيمهم على عدد المعاملات 195 نحصل على 0.07.

قدرت النتائج ب 30 قاعدة. يتم عرض بعض تلك النتائج كالتالي:

	support	itemsets
0	0.441026	(brownie)
1	0.164103	(cinnamon)
2	0.097436	(coconut raisin)
3	0.092308	(cream)
4	0.092308	(dark chocolate)
5	0.071795	(dried fruit)
6	0.123077	(full popcorn)
7	0.266667	(half popcorn)
8	0.071795	(kinder)
9	0.179487	(lotus)
10	0.174359	(milka)

15	0.076923	(walnut)
16	0.082051	(brownie, cinnamon)
17	0.092308	(brownie, half popcorn)
18	0.082051	(brownie, lotus)
19	0.082051	(brownie, milka)
20	0.179487	(brownie, nutella)
21	0.107692	(brownie, smores)
22	0.271795	(brownie, sweet chocolate)
23	0.071795	(sweet chocolate, cinnamon)
24	0.076923	(lotus, coconut raisin)
25	0.076923	(sweet chocolate, half popcorn)
26	0.092308	(milka, nutella)

لعرض القواعد الارتباط بنوعين أو أكثر يجب علينا أن نكتب الرماز المصدري التالي:

```
rules = association_rules(frequent_itemsets, metric="lift", min_threshold=1.2)
rules.sort_values('lift', ascending=False)
```

مع إبقاء الMin_Support في ترميز الأخير، تم إضافة ال Lift أو الرفع والذي يحدد نسبة اعتماد الأنواع ببعضها. Min_Threshold = 1 أي الأنواع غير معتمدة على بعضها تماما. لذلك تم اختيرت النسبة أن تكون 1.2 أي معتمد

تظهر لدينا 15 قاعدة التالية:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
6	(lotus)	(coconut raisin)	0.179487	0.097436	0.076923	0.428571	4.398496	0.059435	1.579487
7	(coconut raisin)	(lotus)	0.097436	0.179487	0.076923	0.789474	4.398496	0.059435	3.897436
13	(nutella, sweet chocolate)	(brownie)	0.138462	0.441026	0.102564	0.740741	1.679587	0.041499	2.156044
14	(brownie)	(nutella, sweet chocolate)	0.441026	0.138462	0.102564	0.232558	1.679587	0.041499	1.122611
4	(brownie)	(sweet chocolate)	0.441026	0.374359	0.271795	0.616279	1.646225	0.106693	1.630458
5	(sweet chocolate)	(brownie)	0.374359	0.441026	0.271795	0.726027	1.646225	0.106693	2.040256
8	(nutella)	(milka)	0.333333	0.174359	0.092308	0.276923	1.588235	0.034188	1.141844
9	(milka)	(nutella)	0.174359	0.333333	0.092308	0.529412	1.588235	0.034188	1.416667
12	(nutella, brownie)	(sweet chocolate)	0.179487	0.374359	0.102564	0.571429	1.526419	0.035371	1.459829
15	(sweet chocolate)	(nutella, brownie)	0.374359	0.179487	0.102564	0.273973	1.526419	0.035371	1.130140
2	(brownie)	(smores)	0.441026	0.164103	0.107692	0.244186	1.488009	0.035319	1.105957
3	(smores)	(brownie)	0.164103	0.441026	0.107692	0.656250	1.488009	0.035319	1.626107
11	(milka)	(sweet chocolate)	0.174359	0.374359	0.082051	0.470588	1.257051	0.016778	1.181766
10	(sweet chocolate)	(milka)	0.374359	0.174359	0.082051	0.219178	1.257051	0.016778	1.057400
0	(nutella)	(brownie)	0.333333	0.441026	0.179487	0.538462	1.220930	0.032479	1.211111
1	(brownie)	(nutella)	0.441026	0.333333	0.179487	0.406977	1.220930	0.032479	1.124183

كما نرى، تم ترتيب القواعد تنازليا حسب أعلى Lift. في الجدول، يوجد سابق Antecedent ولاحق consequent، يعبران عن علاقة طرفا القواعد ببعضهما. أي شراء مادة أ يؤدي إلى شراء مادة ب وهكذا. تعبر Leverage عن الفرق بين التكرر المرصود ل أ وب بظهورهما معا والتكرر المتوقع إذا كان أ وب مستقلين. (0)تعني أن أ وب مستقلين تماما. تعبر قيمة Conviction عن اعتماد العنصر اللاحق Consequent بالعنصر السابق Antecedent. إذا نظرنا إلى الجدول أعلاه، على القاعدة الثانية رقمها 7، نرى أن Lotus معتمد بشكل كبير على Coconut raisin بقيمة 3.89، وهي قيمة عالية بالنسبة لباقي القواعد. أي أنّه الأكثر اعتمادا على سابقه من بين كل القواعد الثانية.

اختبار احتمالات

الآن لنفترض أن لدينا زبون يحب نوع Lotus كثيرا ويشتريه باستمرار، والآن يريد أن يشتري ذلك النوع مجددا. من خلال نظام التوصية الذي لدينا، ومجموعة البيانات التي لدينا، من الممكن أن نكتب الرماز المصدري التالي الذي سيرشح لنا كل اللواحق بالنوع Lotus.

```
rules[ rules['antecedents'].str.contains('lotus', regex=False) &
      rules['consequents'] == 1].sort_values("lift", ascending=False)
```

لاحظ أن المقاييس التي استخدمت سابقا (lift)، هي نفسها تستخدم في الرماز المصدري أعلاه.

والنتيجة هي كالتالي:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
6	(lotus)	(coconut raisin)	0.179487	0.097436	0.076923	0.428571	4.398496	0.059435	1.579487

تم العثور على نتيجة واحدة فقط، كما هو مبان في الجدول.

لنفترض الآن أنه يوجد زبون آخر لا يحب نوع Brownie ويريد شراء أنواع عدة. يمكننا من خلال كتابة رماز مصدري جديد، أن نحذف الBrownie من السوابق واللواحق معا ليتم ترشيح كل القواعد بحسب المقاييس المكتوبة سابقا.

```
rules[~rules['antecedents'].str.contains('brownie', regex=False) &
      ~rules['consequents'].str.contains('brownie', regex=False)].sort_values("lift", ascending=False)
```

وتكون النتائج هي كالتالي:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
6	(lotus)	(coconut raisin)	0.179487	0.097436	0.076923	0.428571	4.398496	0.059435	1.579487
7	(coconut raisin)	(lotus)	0.097436	0.179487	0.076923	0.789474	4.398496	0.059435	3.897436
8	(nutella)	(milka)	0.333333	0.174359	0.092308	0.276923	1.588235	0.034188	1.141844
9	(milka)	(nutella)	0.174359	0.333333	0.092308	0.529412	1.588235	0.034188	1.416667
11	(milka)	(sweet chocolate)	0.174359	0.374359	0.082051	0.470588	1.257051	0.016778	1.181766
10	(sweet chocolate)	(milka)	0.374359	0.174359	0.082051	0.219178	1.257051	0.016778	1.057400

يمكننا أن نفترض الن أنه يوجد زبون قد طلب منتجين، على سبيل المثال Nutella و brownie.

يمكننا أن نكتب الرماز المصدري التالي، حيث نحدد بأن تكون السوابق هي المنتجين السابقين.

```
rules[ rules['antecedents'].str.contains('brownie', regex=False) & rules['antecedents'].str.contains('nutella', regex=False) &
      rules['consequents'] == 1].sort_values("lift", ascending=False)
```

فتكون النتيجة هي كالتالي:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
13	(nutella, brownie)	(sweet chocolate)	0.179487	0.374359	0.102564	0.571429	1.526419	0.035371	1.459829

كما نرى فإن عند طلب المنتجين Nutella وBrownie معا، اللاحق لهم هو Sweet chocolate فقط. بالطبع من الممكن كتابة شرط لكل منهم لوحده كما في رماز الlotus سابقا، وقد تكون النتائج مختلفة وأشمل للمنتجين. ولكن للحصول عليهم في جدول واحد نكتب الرماز المصدري التالي:

```
rules[ rules['antecedents'].str.contains('brownie', regex=False) | rules['antecedents'].str.contains('nutella', regex=False) &
      rules['consequents'] == 1].sort_values("lift", ascending=False)
```

وتكون النتيجة كالتالي:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
12	(sweet chocolate, nutella)	(brownie)	0.138462	0.441026	0.102564	0.740741	1.679587	0.041499	2.156044
15	(brownie)	(sweet chocolate, nutella)	0.441026	0.138462	0.102564	0.232558	1.679587	0.041499	1.122611
5	(brownie)	(sweet chocolate)	0.441026	0.374359	0.271795	0.616279	1.646225	0.106693	1.630458
8	(nutella)	(milka)	0.333333	0.174359	0.092308	0.276923	1.588235	0.034188	1.141844
13	(nutella, brownie)	(sweet chocolate)	0.179487	0.374359	0.102564	0.571429	1.526419	0.035371	1.459829
2	(brownie)	(smores)	0.441026	0.164103	0.107692	0.244186	1.488009	0.035319	1.105957
0	(nutella)	(brownie)	0.333333	0.441026	0.179487	0.538462	1.220930	0.032479	1.211111
1	(brownie)	(nutella)	0.441026	0.333333	0.179487	0.406977	1.220930	0.032479	1.124183

كما نرى، فإن في عمود السوابق يوجد إما المنتج Nutellaأو المنتج Brownie، والنتائج اختلفت وأصبحت الخيارات أكثر.

الخاتمة والآفاق المستقبلية

ختاما، يمكن الاستنتاج أنه من خلال تطبيق خوارزمية apriori، يوفر نظام التوصية المنتجات لعملاء المحل عبر الإنترنت بناءً على تكرار المشتريات وارتباطها ببعضها. يتمثل تطبيق طريقة Apriori في هذا البحث في العثور على معظم مجموعة العناصر بناءً على بيانات المعاملة ثم تكوين نمط ارتباط بين مجموعة العناصر.

بسبب البيانات المحدودة التي لدينا، فالنظام يفتقر الى الدقة العالية. لذلك مستقبلاً، نريد تجريب النظام على المحل لاختبار دقته ومن ثم تحسينه كمتابعة، من الأجل الحصول على أداء أفضل وبالتالي تقديم توصيات أكثر دقة من خلال اتباع عدة طرق:

1. توسيع قاعدة البيانات وإضافة عوامل أخرى، على سبيل المثال، التقييم وإضافة وزن لكل عميل وإضافة خصائص للمنتجات.
2. اختبار نظم توصية بالاستعانة بالإضافات السابقة. كنظام التوصية المعتمد على الترشيح التعاوني والذي يعتمد على تقييمات العميل للمنتج.

مصادر:

1- G. Piatetsky-Shapiro and W.J. Frawley. Knowledge Discovery in Databases. 1991.
2- Jiawe Han, Micheline Kamber, Jian Pei. Data mining: Concepts and Techniques ,3rd ed. (Morgan Kaufmann publisher,2011).
3- CHOUDHARY, A.K., HARDING, J.A. and TIWARI, M.K., 2009. Data mining in manufacturing: a review based on the kind of knowledge
4- Data Mining and Business Analytics with R, First Edition. Johannes Ledolter.(2013 John Wiley & Sons, Inc.)
5- Radwan Mohammed. Data Mining, 2014.
6- K. Ming Leung. Naïve Bayesian Classifier, 2007.
7- Xiao-Li Li and Bing Liu. Rule-Based Classification, 2014.
8- Jeremy Arkes. Regression Analysis: A Practical Introduction, 2019.
9- Swati Gupta. A Regression Modeling Technique on Data Mining, 2015.
10- Beyer, W. H. CRC Standard Mathematical Tables and Formulae, 31st ed. 2002
11- Daniel Peña, George C. Tiao, Ruey S. Tsay. A Course in Time Series Analysis. 2001
12- Christopher Chatfield. The Analysis of Time Series: Theory and Practice. 1975
13- Mohini Chakarverti, Nikhil Sharma and Rajiva Ranjan Divivedi. Prediction Analysis Techniques of Data Mining: A Review. 2019

14- J. Wolberg. Prediction Analysis in Designing Quantitative Experiments. 2010

15- Mahamed G.H. Omran, Andries P. Engelbrecht and Ayed Salman. An overview of clustering methods. 2007

16- Preeti Arora, Dr. Deepali, Shipra Varshney. Analysis of K-Means and K-Medoids Algorithm for Big Data. 2015

17- Mohiuddin Ahmed, Abdum Naser Mahmud. Clustering Based Semantic Data Summarization Technique: A New Approach. 2014

18- Hegland, M. Algorithms for Association Rules. 2003

19- Ming-Yen Lin, Pei-Yu Lee, Sue-Chen Hsueh. Apriori-based Frequent Itemset Mining Algorithms on MapReduce. 2012

20- J Chandra, K R Dewi. Implementation of Data Mining Sales of Milk Using Apriori Algorithm Method. 2019

21- Arijita Das, Soumita Jana, Pranita Gangouly. Application of Association Rule:Apriori Algorithm in E-Commerce. 2021

22- سميرة محمد علي القدم. تطبيق تقنيات التنقيب في البيانات لتقيم أداء طلاب قسم الحاسوب. 2018

23- Primasatria Edastama, Ankur Singh Bist, Ari Prambudi. Implementation of Data Mining on Glasses Sales Using the Apriori Algorithm. 2021

24- Zhuang Chen, Shibang Cai, Qiulin Song and Chonglai Zhu. An Improved Apriori Algorithm Based on Pruning Optimization and Transaction Reduction. 2011

25- Santosh Shakya, Anju Singh, Divakar Singh. A Survey on Hash based A-priori Algorithm for Web Log Analysis. 2013

26- Francesco Ricci, Lior Rokach, Bracha Shapira. Recommender Systems Handbook. 2015

27- Nour Nassar. Recommender System in Big Data. 2019

28- Selma Afif, Zaki Brahmi, Mohamed Mohsen Gammoudi. Trust-Based Recommender System: An Overview. 2016