مجلة جامعة طرطوس للبحوث والدراسات العلمية _ سلسلة العلوم الهندسية المجلد (6) العدد(7) 2022
Tartous University Journal for Research and Scientific Studies - engineering Sciences Series Vol. (6) No. (7) 2022

استخدام ومقارنة خوارزميات التنقيب عن البيانات في التسويق

م. علي هيثم محمد*

(تاريخ الإيداع 2/3 /2022 . قُبِل للنشر في 2022/7/31)

□ ملخّص□

مع توسع الاسواق الاقتصادية وتنوعها وتطورها المستمر المترافق مع التطور الكبير والسريع في المجالات التقنية وعالم الانترنت ظهرت الكثير من التقنيات لمساعدة الشركات التجارية في تلبيتها لحاجة هذه الأسواق فتطرقنا في هذه الدراسة إلى أهم الأشياء التي تقوم بها الشركات الضخمة عالمياً من أجل المحافظة على زبائنها وإرضائهم مما يزيد من أسهمها وأرباحها و تحدثنا عن التنبؤ بالأسواق التجارية من خلال التنقيب عن البيانات الذي سيفيد هذه الشركات في الكثير من العمليات التجارية التي ستقوم بها حيث سيعطيها فكرة كاملة عن واقع السوق وعن واقع الزبون مما يساعدها في تقديم العروض مثلاً أو تحديد سعر السلعة حسب متطلبات المكان الذي ستبيعها فيه والزمان المناسب لرفع وخفض أسعار بضائعها بالإضافة لكثير من الميزات التي ممكن أن تقدمها مثلاً لزبائنها المميزين في الوقت المناسب وقمنا بقياس دقة بعض الخوارزميات المتبعة في هكذا مجالات .

تم في هذا البحث دراسة بعض الفروق بين خوارزميات التنقيب في البيانات المتعددة والتي من خلالها يمكن مناقشة أفضل الطرق التي يمكن استخدامها من أجل التأكد من أن المنتجات التي ستعرض للزبون هي المنتجات التي فعلاً يهتم بها من خلال دمج خوارزميات التنقيب في البيانات واستخدامها مع بعضها لتحقيق هذا الغرض

كلمات مفتاحية: التنقيب عن البيانات – السلة التجارية – النظام الناصح

* م. علي هيثم محمد – ماجستير في علوم الانترنت

مجلة جامعة طرطوس للبحوث والدراسات العلمية _ سلسلة العلوم الهندسية المجلد (6) العدد(7) 2022
Tartous University Journal for Research and Scientific Studies - engineering Sciences Series Vol. (6) No. (7) 2022

Use and comparison of data mining algorithms in marketing

(Received 3/2/ 2022 . Accepted 31/7/ 2022)

□ ABSTRACT

With the expansion, diversity, and ongoing evolution of markets, along with the immense and rapid development of technical fields and the Internet, many techniques have appeared to assist commercial companies in meeting the needs of these markets. This study approached the most significant measures big companies carry out globally to maintain their costumers and satisfy their needs. In doing so, they raise the value of their shares, along with their profits. The study talked about market predicting by means of data mining. This helps companies throughout many of their business processes by providing a whole idea about market as well as costumer realities. This also helps companies in providing offers, pricing according to the place of selling, and in defining the right time to raise and lower prices of their goods, in addition to many advantages they could present to their special costumers in the right time. The study measured the accuracy of some followed algorithms in such fields.

This research studied some of the differences amid algorithms of multi-data mining. This would make possible the discussion of the best methods to be utilized in order to confirm that the products, which are to be shown to the customer, are the same products the customer is interested in. This is done through the integration and employment of the various algorithms of data mining all together.

Key words: Data Mining, Market Basket, Recommender System.

* Researcher holds a master's degree (Internet)

1- مقدمة

مع الازدياد الكبير لمفهوم التسويق عبر الانترنت أصبحت أغلب الشركات تحاول فهم ما يطلبه الزبائن من أجل الحصول على رأي الزبائن ومحاولة إرضاء الزبائن بعرض المنتجات التي يرغب بها الزبون أولاً وذلك بناءً على الكثير من المعطيات والبيانات للوصول إلى مرحلة المعرفة ضمن هذا المجال بحيث تكون الشركات قادرة على التنبؤ بما يريده المستخدم وما يحتاجه من منتجات وذلك بناءً على معلومات سابقة أو حتى مشتريات الزبون السابقة.

إن التطور الكبير والسريع لمفهوم التنقيب في البيانات وخاصة مفهوم السلة التجارية والتي يمكن من خلالها التنبؤ بمجموعة المنتجات التي قد يطلبها الزبون مع بعضها أصبح من أهم المفاهيم المتبعة في عالم التسويق والتجارة الالكترونية بحيث تظهر للزبون المنتجات التي تهمه بناءً على اختياره لمنتج واحد، ثم ظهرت مفاهيم أخرى كان أهمها عرض المنتجات التي تهم الزبون بناءً على صفاته وبناءً عليها أصبحت الشركات تحاول الوصول إلى صفات الزبائن بحيث تقوم بوضع الزبائن ضمن عناقيد يمكن من خلالها التنبؤ بما سيطلبه المستخدم بناءً على صفاته.

2-هدف البحث

يهدف هذا البحث إلى:

- دراسة تقنيات التنبؤ المؤتمت من خلال التنقيب عن البيانات في مجال التسويق وتنظيم العلاقات التجارية مع الزبائن .
- تأكيد أهمية تقنيات التنبؤ في العمليات التجارية .
- تبيان محاسن هذه التقنيات وتأثيرها الجيد على الشركات التي تسعى إلى إرضاء زبائنها .
- مقارنة بعض خوارزميات التنبؤ المتبعة , وحساب دقتها و موثقيتها .

3-مشكلة البحث:

مع توسع الأسواق التجارية وتطور الشركات والحاجة الماسة لتقنيات التنقيب عن البيانات وضرورة استخدامها في الأسواق التجارية وجدنا أنه من الضروري معرفة مزايا هذه التقنيات وضرورة فهم الية عملها وكيفية حساب دقة هذه الخوارزميات كي تكون موثوقة من قبل الشركات التجارية

4-منهجية البحث وطرائقه

قمنا في هذا البحث بدراسة عن تقنيات التنقيب عن البيانات وارتباطها بالتنبؤ بالأسواق التجارية وفهم طرق وأساليب هذه التقنيات والخوارزميات ومقارنتها

حيث اعتمد هذا البحث في تنفيذه على العديد من المراجع وعلى التطبيق العملي لمقارنة عمل خوارزميات و تقنيات التنقيب عن البيانات على موقع قمنا بتصميمه وتنفيذ هذه الخوارزميات عليه ودراسة نتائج التنفيذ

كما اعتمدنا في الدراسة العملية على برنامج Microsoft Visual Studio وهو بيئة تطوير متكاملة تم تصميمه لتطوير برامج الكمبيوتر ومواقع الانترنت واستخدمنا SQL Server كنظام لإدارة قاعدة البيانات

5-التنقيب في البيانات

تهدف تقنية التنقيب في البيانات لاستخلاص المعلومات ضمن كمية كبيرة من البيانات، وتحويلها إلى صيغة قابلة للفهم بهدف استخدامها في أمور أخرى، أي أن الهدف الرئيسي هو تحديد البيانات القيمة المفيدة بشكل جزئي القابلة للفهم المترابطة ذات الصيغة النمطية ضمن حجم هائل للبيانات

5-1-السلة التجارية Market Basket:

يعتبر تحليل السلة التجارية MBA تقنية نمذجة خاصة بالتجارة الإلكترونية يساعد في تحديد مجموعة العناصر التي يمكن شراءها لنفس الغرض من عملية الشراء، لذلك يعرف أيضاً بتحليل التقارب[1].

بدأ العمل بهذا مصطلح ما قبل التجارة الإلكترونية ضمن المحلات التجارية بهدف توضيع العناصر التي يمكن أن يطلبها الزبون سويةً في مكان واحد، وأصبح بعد استخدام التجارة الإلكترونية يشكل سلسلة مترابطة من المواد – مجموعة عناصر- يتم تقديمها للزبون بمجرد طلب أحد عناصر السلسلة وهنا دخل مصطلح آخر يصب في نفس المعنى وهو قواعد الربط Association Rules [1].

تتطلب عملية بناء هذه المجموعات من العناصر تقنية التنقيب في البيانات للبحث ضمن كمية كبيرة من البيانات التي تتضمن عمليات شراء للزبائن ومجموعة العناصر التي تم شراؤها، بالإضافة إلى تقنيات التوقع لتحديد الخيار الأنسب الذي يتم تقديمه للزبون.

5-1-1-تحليل السلة التجارية

تستخدم محلات السوبر ماركت البيانات لتحسين فهم احتياجات المتسوقين و زيادة قيمة الإنفاق العام، وتعتبر تقنية تحليل السوق التجارية MBA واحدة من التقنيات الرئيسية المستخدمة من قبل التجار فهي تكشف عن العلاقة بين المنتجات من خلال البحث عن مزيج من المنتجات التي تباع معاً بشكل متكرر في المعاملات مما يسمح للمحلات بالتعرف على العلاقات بين المنتجات التي يشتريها الأفراد [12].

على سبيل المثال ، من المحتمل أن العملاء الذين يشترون قلم رصاص وورقة يقومون بشراء ممحاة أو مسطرة.يتيح تحليل سلة السوق لتجار التجزئة تحديد العلاقات بين المنتجات التي يشتريها الناس [9].

5-2-النظام الناصح Recommender System:

وهو عبارة عن منظومة خاصة بتحليل البيانات الخاصة بالزبون وتطبيق آلية توقع عليها بهدف تقديم لائحة تذكير بكل البيانات التي تحقق ارتباط وثيق قد يحتاجها الزبون، وتعتبر التجارة الالكترونية من أكثر الأمثلة وضوحاً واستخداماً لهذه التقنية[2].

وتصنف النظم الناصحة وفقاً لآلية عملها إلى:

1. نظام ناصح عام General.
2. نظام ناصح تتابعي Sequential.
3. نظام ناصح نمطي Pattern-Based.

4. نظام ناصح هجين Hybrid.

النظام الناصح	تتابع المعلومات	الزبائن الجدد	استقلالية العناصر	عامل التأثير المتبادل	وجود سابقة	تفاعلات
العام	-	-	✓	✓	-	
التتابعي	✓	✓	-	-	-	
النمطي	-	-	✓	✓	-	
الهجين	✓	✓	✓	✓	-	

جدول : 1تصنيف النظم الناصحة

5-3-خوارزمية التصنيف

كما بات معروفا فإن خوارزميات التصنيف هي شكل من أشكال تحليل البيانات والتي تستخلص نماذج تصف بشكل دقيق فئات وتصنيفات البيانات المهمة.

وتستخدم خوارزميات التصنيف في العديد من المجالات ولكثير من التطبيقات مثل كشف عمليات الاحتيال، والتسويق المستهدف لفئات معينة والتنبؤ بمستوى الأداء أو مدى الإقبال على شراء المنتجات، وتشخيص الأمراض. فهي تنتمي إلى النموذج التنبؤي الخاضع للإشراف من نماذج التنقيب في البيانات[11].

5-3-1-أنواع خوارزمية التصنيف

- التصنيف باستخدام خورازميات شجرة القرار.
- التصنيف باستخدام خوارزميات الشبكات العصبية.
- التصنيف باستخدام نظرية الاحتمالات.
- التصنيف باستخدام خوارزمية الجار الأقرب.

5-4-مفهوم التحليل العنقودي

يختلف التحليل العنقودي عن خوارزميات التصنيف فالمجموعة أو الفئة التي يمكن أن ينتمي لها الزبون هنا غير معروفة، كما أنه بوجود عدد كبير من الزبائن وتعدد السمات التي تصفهم يصعب أو ربما مستحيل تجزئتهم يدوياً لمجموعات تخدم هذا الهدف، وهنا تظهر الحاجة لاستخدام تقنية التحليل العنقودي Clustering.

إن خوارزميات التحليل العنقودي تتضمن عملية تجميع مجموعة من البيانات بداخل عدد من الفئات أو الأجزاء، بحيث تضم كل مجموعة أو فئة العناصر الأكثر تشابهاً ولكنهم أكثر اختلافاً عن العناصر التي تنتمي للمجموعات الأخرى[11].

كما يتم تقييم التشابه وعدم التشابه بناءً على قيم السمات التي تصف البيانات، والتي غالباً ما تشتمل على قياس المسافات فيما بينها

أن التحليل العنقودي كأحد أدوات التنقيب في البيانات له تطبيقات في مجالات عديدة سواء في مجال الطب والعلوم البيولوجية والمعلوماتية الحيوية واستخبارات الأعمال أو ذكاء الأعمال وبحوث الشبكة العنكبوتية والمجالات الأمنية وغيرها من المجالات.

5-4-1-أساليب التحليل العنقودي

التحليل العنقودي أو التحليل بالتجزئة العنقودية هي عملية تجزئة مجموعة من البيانات إلى مجموعات جزئية، وكل مجموعة جزئية تمثل كتلة Cluster من البيانات بحث يكون عناصر كل مجموعة متشابهة وبنفس الوقت مختلفة عن بقية العناصر في المجموعات الأخرى، وضمن هذا السياق يمكن لطريقتين من طرق التجزئة أن ينتج عنهما مجموعات جزئية مختلفة لنفس البيانات، وتتم عملية التجزئة باستخدام خوارزميات تقود إلى توصيف المجموعات الجزئية التي لم تكن معروفة قبل التجزئة[11].

وكأحد أدوات تنقيب البيانات فإن خوارزمية التحليل العنقودي يمكن أن تكون أداة تنقيب قائمة بذاتها وتهدف لاكتساب المعرفة المخبأة بداخل قواعد البيانات وملاحظة خصائص كل مجموعة جزئية والتركيز على مجموعات محددة لمزيد من التحليل ولكنها يمكن أن تكون خطوة من خطوات تحضير البيانات للتحليل والتنقيب لكي تخدم غيرها من خوارزميات التنقيب.

ونظراً لازدياد استخدام تقنيات التحليل العنقودي في العديد من المجالات وبخاصة تلك التي تتعامل مع كميات هائلة من البيانات فقد أصبح هذا المسار من المسارات الهامة في بحوث التنقيب في البيانات.

6-التطبيق العملي

اخترنا كتطبيق عملي موقع بيع التكروني (قمنا نحن بتصميمه) لمجموعة منتجات وسنبني خوارزمية توقع لمشتريات الزبون حسب المواصفات الخاصة بالزبون وذلك بناءً على قواعد بيانات قديمة تم الحصول عليها من الانترنت للعمل عليها وتجريب الموقع الذي سنقوم ببنائه

6-1-مراحل العمل

6-1-1-التقنية الأولى

وضمن هذه التقنية سنقوم ب

- عنقدة المنتجات باستخدام مواصفات الزبائن
- استخدام خوارزمية شجرة القرار في الحصول على فئة المنتجات من خلال مواصفات الزبائن
- التحقق من أن الفئة تحتوي على المنتج ووضع true أو false حسب النتيجة
- حساب قيمة الزمن اللازم للقيام بالعمليات السابقة
- يتم استخدام خوارزمية ال clustering باستخدام برنامج Visual Studio من أجل عنقدة المنتجات باستخدام مواصفات الزبون ليستخدم الخرج الخاص بهذه الخوارزمية كدخل من أجل عملية التحقق من أن فئة المنتجات التي قد يختارها الزبون حسب مواصفاته صحيحة أو لا بعد أن يتم استخدام

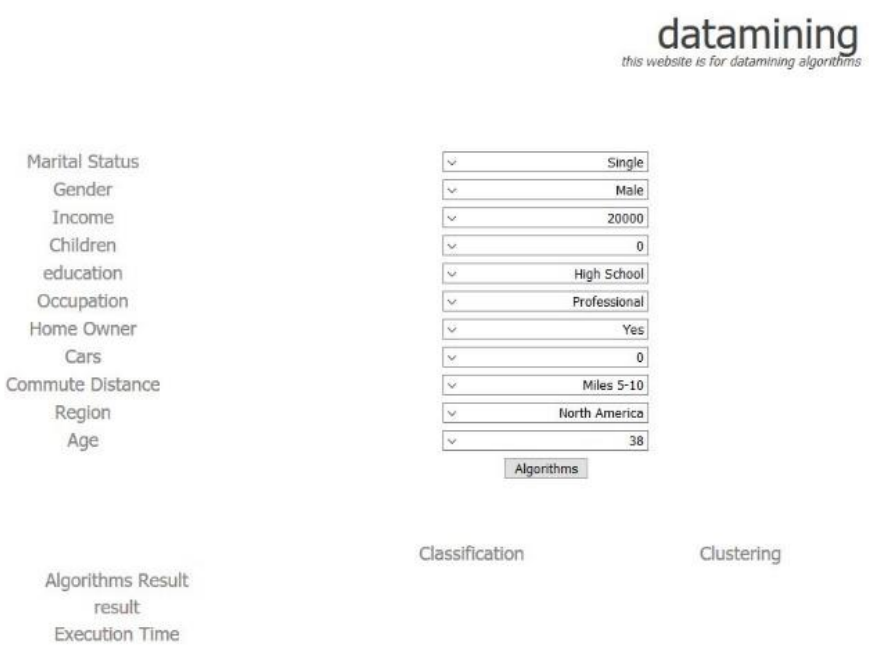
برنامج Visual Studio من أجل إنشاء خوارزمية Decision Tree التي سيتم استخدامها للتصنيف وسيتم تجريب مجموعة من بيانات الزبائن ومدى اقترابها من النتائج الصحيحة وسيتم وضع النتائج ضمن جدول موضح فيه كل عملية من العمليات التي تم استخدامها ليتم استخدام المقاييس الخاصة بخوارزمية ال Classification لقياس دقة الخوارزمية ومن خلال صفحات الويب أيضاً سيتم احتساب الزمن الخاص بتنفيذ الخوارزمية ووضع النتائج في جدول منفصل ليتم بعدها حساب متوسط الوقت المطلوب لتنفيذ الخوارزمية.

6-1-2-التقنية الثانية

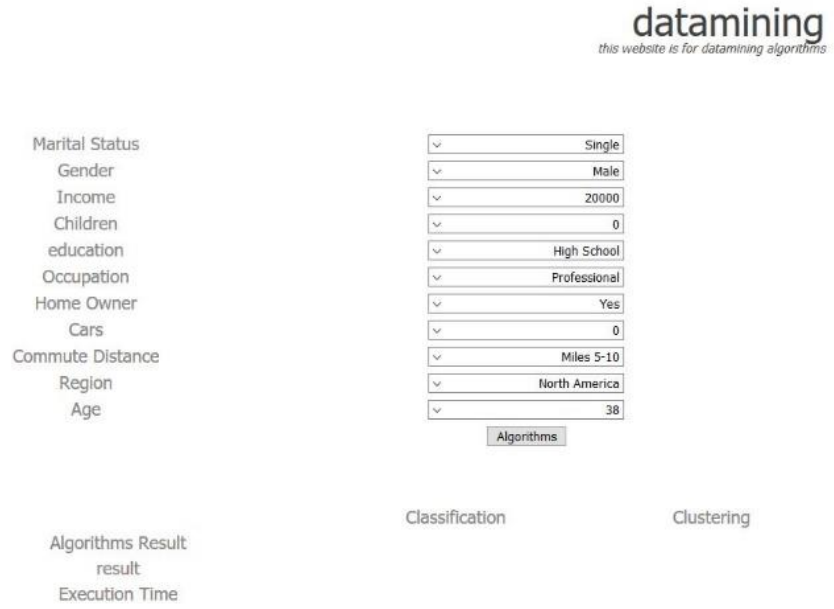
- عنقدة الفئات باستخدام مواصفات الزبائن
- استخدام خوارزمية شجرة القرار في الحصول على المنتجات باستخدام مواصفات الزبائن
- التحقق من أن المنتج يتبع لفئة منتجات
- وضع القيمة true أو false حسب نتيجة الخطوة السابقة
- يتم بناء خوارزمية ال Clustering باستخدام برنامج Visual Studio من أجل عنقدة فئات المنتجات حسب مواصفات الزبون ومن ثم بناء خوارزمية ال Decision Tree من أجل تصنيف المنتجات حسب مواصفات الزبون ليتم التحقق من أن الزبون صاحب مجموعة المواصفات المحددة سيقوم بشراء منتجات من نفس الفئة وسيتم إجراء مجموعة من الاختبارات على مجموعة من مواصفات الزبائن ليتم بعدها وضع جميع النتائج ضمن جدول مفصل ويتم استخدام النتائج في هذا الجدول ليتم استخدام بعدها مقاييس الدقة الخاصة بخوارزمية ال Classification للتحقق من دقة الخوارزمية ومن خلال صفحات الويب سيتم احتساب زمن تنفيذ الخوارزمية ووضعها ضمن جدول منفصل ليتم بعدها حساب متوسط الوقت المطلوب لتنفيذ الخوارزمية.
- ومن ثم ستتم مقارنة نتائج الطريقتين في استخدام الخوارزمية من خلال مقاييس الدقة ومتوسط زمن تنفيذ كل طريقة من الطرق المذكورة أعلاه ليتم بعدها استخلاص النتائج واقتراح التوصيات.

6-2-الموقع الالكتروني

سنقوم بإدخال البيانات عن طرق موقع بسيط قمنا بتصميمه من اجل هذا الغرض وذلك من خلال الصفحتين في الشكلين التاليين:

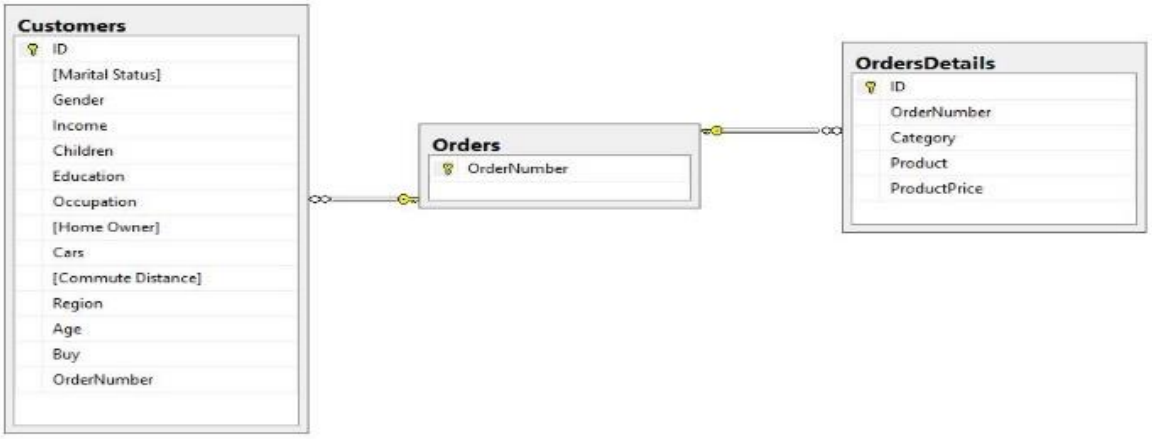


شكل : 1عنقدة المنتجات وتصنيف الفئات



شكل : 2عنقدة الفئات وتصنيف المنتجات

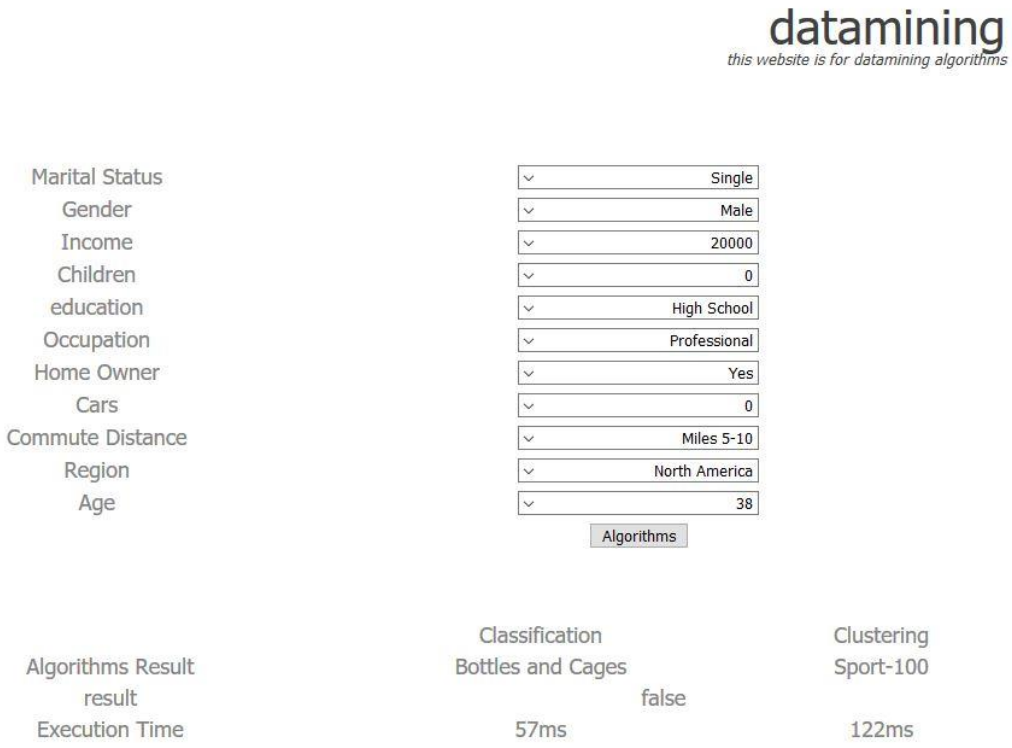
تم الحصول على البيانات من قاعدة بيانات خاصة بالحركة التجارية وهي على الشكل التالي:



شكل : 3قاعدة البيانات المستخدمة

6-3-بناء الخوارزميات

تم بناء خوارزمية العنقدة وخوارزمية التصنيف باستخدام برنامج Visual Studio ومن خلال استخدام نفس قاعدة البيانات الموضحة بالشكل 3وبعد التنفيذ كانت النتائج كالتالي:



شكل : 4نتيجة التنفيذ

6-4-التقنية الأولى:

تم اختبار الموقع على 75 قيمة مختلفة لخصائص الزبائن وكانت النتائج على الشكل التالي:

marital status	gender	income	children	Education	Occupation	home owner	cars	commute distance	Region	Age	predict result	real result
Single	male	20000	0	high school	Professional	Yes	0	mailes 5-10	north america	30	FALSE	FALSE
married	female	20000	4	high school	Professional	Yes	0	mailes 5-10	north america	30	FALSE	FALSE
married	female	20000	4	partial collage	Clerical	Yes	0	mailes 5-10	north america	30	FALSE	TRUE
married	female	20000	4	partial collage	Clerical	No	0	mailes 2-5	north america	30	TRUE	TRUE
married	female	20000	4	partial high collage	Clerical	No	0	mailes 2-5	north america	47	TRUE	TRUE
married	female	20000	4	graduate degree	Manual	No	0	mailes 2-5	north america	47	FALSE	TRUE
Single	male	90000	3	Bachelors	Management	Yes	4	mailes 1-2	Europe	28	FALSE	FALSE
Single	male	90000	3	Bachelors	skilled manual	Yes	4	mailes 1-2	Europe	28	FALSE	FALSE
Single	male	90000	3	high school	skilled manual	No	4	mailes 1-2	Europe	28	TRUE	TRUE
Single	male	120000	1	graduate degree	Clerical	No	4	mailes 1-2	Pacific	32	FALSE	FALSE
Married	female	120000	1	graduate degree	Clerical	No	4	mailes 2-5	Pacific	44	TRUE	TRUE
Married	female	120000	1	graduate degree	Clerical	No	4	mailes 2-5	north america	44	FALSE	FALSE
Married	female	60000	2	high school	Professional	Yes	3	mailes 0-1	Europe	48	TRUE	TRUE
Single	male	60000	2	partial collage	Manual	Yes	3	mailes 0-1	Europe	48	FALSE	FALSE
Married	male	60000	2	partial high collage	Clerical	No	3	mailes 0-1	Pacific	33	FALSE	FALSE
Married	male	100000	2	partial high school	Manual	No	2	mailes +10	north america	33	FALSE	TRUE
Single	male	100000	2	graduate degree	Professional	No	2	mailes +10	north america	37	FALSE	TRUE
Single	male	100000	2	Bachelors	skilled manual	No	2	mailes +10	Pacific	26	TRUE	TRUE
Single	male	50000	2	Bachelors	skilled manual	No	2	mailes 5-10	Europe	26	TRUE	TRUE
Married	female	110000	0	high school	Professional	Yes	0	mailes 2-5	north america	35	TRUE	TRUE
Married	female	110000	0	partial high school	Clerical	yes	0	mailes 2-5	north america	35	TRUE	FALSE
Married	female	110000	0	partial high school	Clerical	yes	3	mailes 1-2	Pacific	41	FALSE	FALSE
Single	male	70000	3	partial high school	Clerical	yes	3	mailes 1-2	Pacific	41	FALSE	FALSE
Single	male	70000	3	partial collage	Manual	yes	3	mailes 1-2	Pacific	41	FALSE	FALSE
Single	male	80000	4	graduate degree	Management	yes	3	mailes 1-2	Pacific	41	TRUE	FALSE
Single	male	80000	4	graduate degree	Management	yes	4	mailes 0-1	Europe	25	TRUE	TRUE
Single	male	80000	4	Bachelors	skilled manual	yes	4	mailes 0-1	Europe	25	FALSE	TRUE
Married	female	110000	1	partial high school	Professional	no	2	mailes +10	north america	39	FALSE	FALSE
Married	female	110000	1	partial collage	Professional	no	2	mailes +10	north america	39	FALSE	FALSE
Married	female	110000	1	graduate degree	Professional	no	2	mailes +10	north america	38	TRUE	TRUE
single	male	20000	5	graduate degree	Manual	no	2	mailes +10	north america	45	TRUE	TRUE
single	male	20000	5	Bachelors	Professional	yes	3	mailes 5-10	north america	45	FALSE	FALSE
married	male	100000	0	partial collage	Management	yes	0	mailes +10	Pacific	40	FALSE	FALSE
single	male	170000	0	partial high school	Professional	no	0	mailes 2-5	Europe	25	TRUE	TRUE
single	male	150000	0	partial high school	Manual	no	0	mailes 2-5	Europe	35	TRUE	FALSE
married	female	150000	0	graduate degree	Manual	no	0	mailes 2-5	Europe	35	FALSE	FALSE
single	male	40000	4	partial collage	Management	no	0	mailes 2-5	Europe	35	TRUE	TRUE
married	male	60000	3	graduate degree	Clerical	yes	4	mailes 0-1	north america	26	FALSE	FALSE
married	male	60000	3	partial collage	Management	yes	4	mailes 0-1	north america	26	TRUE	TRUE

married	male	60000	0	Bachelors	Manual	no	0	mailes 0-1	north america	26	FALSE	FALSE
single	female	60000	3	partial high school	Clerical	yes	0	mailes 2-5	Europe	48	TRUE	TRUE
married	male	120000	3	partial high school	Clerical	no	0	mailes 2-5	Pacific	50	TRUE	TRUE
married	male	120000	3	high school	Professional	no	0	mailes 2-5	Pacific	50	TRUE	TRUE
married	male	150000	4	partial collage	Management	no	3	mailes 2-5	north america	70	FALSE	FALSE
married	female	150000	5	partial collage	Management	no	2	mailes 1-2	Europe	70	FALSE	FALSE
single	male	150000	0	partial high school	Manual	no	1	mailes 1-2	Europe	27	TRUE	TRUE
single	male	20000	0	graduate degree	skilled manual	no	0	mailes +10	north america	25	FALSE	FALSE
single	male	160000	0	partial collage	Management	yes	2	mailes 5-10	Europe	27	FALSE	FALSE
single	female	60000	0	high school	Manual	no	1	miles 5-10	Pacific	27	TRUE	TRUE
married	female	60000	2	high school	Manual	yes	0	miles 5-10	Pacific	33	TRUE	TRUE
married	male	10000	2	high school	Manual	yes	1	miles 5-10	Pacific	33	TRUE	TRUE
married	male	170000	2	high school	Manual	yes	4	miles 1-2	Europe	66	TRUE	FALSE
single	female	70000	0	graduate degree	skilled manual	no	0	miles 2-5	Pacific	28	FALSE	FALSE
married	male	70000	0	high school	skilled manual	yes	0	miles 5-10	Europe	30	FALSE	FALSE
married	male	170000	2	Bachelors	skilled manual	yes	4	miles 0-1	Europe	62	FALSE	FALSE
single	male	20000	0	Bachelors	skilled manual	no	0	miles 0-1	north america	42	TRUE	TRUE
single	male	90000	0	Bachelors	skilled manual	no	0	miles 0-1	north america	55	TRUE	TRUE
single	female	90000	0	high school	skilled manual	no	0	miles 0-1	north america	55	FALSE	FALSE
single	female	90000	0	high school	skilled manual	no	0	miles 0-1	Pacific	42	FALSE	FALSE
married	female	90000	1	high school	skilled manual	yes	0	miles 0-1	Pacific	42	TRUE	TRUE
married	female	170000	1	high school	skilled manual	yes	4	miles 5-10	Pacific	43	TRUE	TRUE
married	female	170000	1	partial high school	Professional	yes	4	miles 5-10	Pacific	43	TRUE	TRUE
married	male	170000	4	partial collage	Management	yes	4	miles 1-2	Europe	89	TRUE	TRUE
married	male	170000	0	partial collage	Management	no	1	miles 2-5	Europe	25	TRUE	FALSE
married	male	40000	0	partial collage	Management	no	0	miles 2-5	Pacific	25	TRUE	FALSE
single	female	20000	0	graduate degree	Clerical	no	0	miles +10	north America	31	FALSE	FALSE
single	female	110000	0	partial collage	Clerical	yes	0	miles +10	north America	47	FALSE	FALSE
married	female	110000	5	Bachelors	Manual	no	1	miles 0-1	Europe	32	FALSE	TRUE
single	male	120000	0	Bachelors	Manual	no	3	miles 0-1	Europe	32	TRUE	FALSE
married	male	120000	2	Bachelors	Manual	no	1	miles 0-1	Europe	32	TRUE	TRUE
married	male	120000	2	partial collage	Manual	no	1	miles 0-1	Europe	32	FALSE	FALSE
married	male	120000	2	partial collage	Manual	no	1	miles 0-1	north america	32	FALSE	FALSE
single	female	50000	3	partial high school	Professional	yes	3	miles 5-10	Pacific	45	FALSE	TRUE
married	male	170000	1	partial collage	Management	no	3	miles 5-10	Europe	45	FALSE	FALSE
married	male	170000	1	partial collage	Management	yes	3	miles 5-10	Europe	72	TRUE	TRUE

جدول : 2تنفيذ التقنية الأولى

لقياس دقة الخوارزمية يوجد مجموعة من المعايير التي سوف نطبقها وهي على الشكل التالي:

Accuracy: وهي تستخدم لقياس دقة الخوارزمية.

Precision: وهي تستخدم لقياس إحكام الخوارزمية.

Recall: وهي تستخدم لقياس استرداد الخوارزمية.

F1−score: وهي تستخدم لقياس دقة الاختبارات.

Sensitivity: وهي تقيس حساسية الخوارزمية.

Specificity: وهي تقيس نوعية الخوارزمية.

نسبة الصحة: وهي تقيس نسبة صحة الخوارزمية.

حتى يتم القياس بشكل صحيح يجب استخدام التعابير التالية وهي:

TP: True Positive وهي التي تم التوقع أنها صحيحة وهي فعلا صحيحة في مثالنا (النتيجة المتوقعة true والنتيجة الحقيقية true)

TN: True Negative: وهي التي تم التوقع أنها خاطئة وهي فعلا خاطئة في مثالنا (النتيجة المتوقعة false والنتيجة الحقيقية false)

FP: False Positive: وهي التي تم التوقع بأنه صحيح وهو في الحقيقة خاطئ في مثالنا (النتيجة المتوقعة true وفي الحقيقة false)

FN: False Negative: وهي التي تم التوقع بأنها خطأ وهو في الحقيقة صحيح في مثالنا (النتيجة المتوقعة false ولكن في الحقيقة true)

من خلال الجدول السابق(جدول 2)حصلنا على قيم TN و TP و FN و FP والتي سنستخدمها لقياس دقة الخوارزمية

6−4−1−حساب الدقة Accuracy

وهي تستخدم القانون التالية:

$$\frac{\text{التنبؤات الصحيحة}}{\text{المجموع الكلي}}$$

أي:

$$\frac{TP + TN}{TP + TN + FP + FN}$$

حسب جدول النتائج السابق يكون القانون كالتالي حيث TN هي التي تم التنبؤ بشكل صحيح انها خاطئة و TP هي التي تم التنبؤ بشكل صحيح انها صحيحة أي مجموعهما هو مجموع التنبؤات الصحيحة :

$$\frac{32 + 29}{75} = \frac{61}{75}$$

=0.814

نلاحظ من النتيجة السابقة أن دقة الخوارزمية مرتفعة حيث تعتبر أي خوارزمية تحقق نتيجة دقة أكثر من 0.75 هي خوارزمية ذات دقة جيدة بينما تعتبر الخوارزميات مع نسبة صحة أكبر من 0.9 ذات دقة ممتازة.

حساب حساسية الخوارزمية Sensitivity

وهي تستخدم القانون التالي:

$$\frac{TP}{P} = \frac{32}{36}$$

$$=0.89$$

النوعية الخوارزمية Specificity

والتي يتم حسابها من خلال القانون التالي:

$$\frac{TN}{N} = \frac{29}{39}$$

$$=0.75$$

حساب نسبة الصحة

وهي تستخدم القانون التالي:

Sensitivity * [T/T+N] + specificity * [N / T+N]

=0.89 * [36/75] + 0.75 * [39/75]

=0.89 * 0.48 + 0.75 * 0.52

=0.4272 + 0.39

=0.8172

بالنظر إلى مقاييس النوعية والحساسية ونسبة الصحة نلاحظ أن الخوارزمية تحقق أداء جيد جداً ولكن وبما أن قيمة حساسية الخوارزمية أكبر من قيمة النوعية هذا يعني بأن الخوارزمية قادرة على توقع الشكل الموجب بشكل أكبر أي أنها قادرة على توقع أن المستهلك سيشتري من نفس فئة المنتجات المتوقعة.

حساب الإحكام Precision

وهي تستخدم القانون التالي:

$$\frac{TP}{TP + FP}$$

حسب جدول النتائج السابق يتم التعويض كالتالي:

$$\frac{29}{29 + 6} = \frac{29}{35}$$

$$=0.829$$

وبالتالي فإن إحكام الخوارزمية جيد جداً وهي تقوم بتوقع المنتجات التي سيشتريها الزبون حسب الفئة التي تم التنبؤ بها بشكل صحيح بنسبة 82.9% أي تقريباً 83% أي ومن بين كل مئة زبون هناك 83 زبون سيشترون حسب فئة المنتجات التي تم التنبؤ بها حسب مواصفاتهم.

حساب قيمة الاسترداد Recall

وهي تستخدم القانون التالي:

$$\frac{TP}{TP + FN}$$

حسب جدول النتائج السابق يتم التعويض كالتالي:

$$\frac{29}{29 + 8} = \frac{29}{37}$$

$=0.784$

قيمة الاسترداد لهذه الخوارزمية جيدة ولكنها منخفضة بالمقارنة مع قيمة الإحكام ولذلك ينصح باستخدام قيمة الإحكام للحصول على فئة المنتجات التي يتوقع من الزبون شراؤها.

حساب قيمة دقة الاختبار F1−Score:

ولها القانون الرياضي التالي:

$$2 * \frac{Precision * Recall}{Precision + Recall}$$

حسب النتائج السابق يتم التعويض كالتالي:

$$2 * \frac{0.829 * 0.784}{0.829 + 0.784} = \frac{1.293}{1.613}$$

$=0.802$

تظهر القيمة السابقة أن دقة الاختبار في مستوى جيد بالمقارنة مع عدد الاختبارات التي قمنا بها حيث سيرتفع هذا العدد مع ارتفاع عدد الاختبارات.

كما أن وقت تنفيذ الخوارزمية هو مقياس مهم في حساب جودة الخوارزمية واستخدامها ومن خلال الاختبارات السابقة يبين الجدول التالية وقت تنفيذ الخوارزميات في ال 75 اختبار:

classification time	clustering time
12ms	165ms
15ms	4ms
13ms	5ms
14ms	5ms
12ms	5ms
16ms	4ms
20ms	10ms
19ms	8ms
16ms	6ms
12ms	4ms
15ms	4ms

13ms	5ms
14ms	5ms
12ms	5ms
16ms	4ms
20ms	10ms
9ms	3ms
26ms	5ms
12ms	7ms
13ms	5ms
12ms	4ms
15ms	4ms
13ms	5ms
14ms	5ms
12ms	5ms
16ms	4ms
20ms	10ms
19ms	8ms
16ms	6ms
12ms	4ms
15ms	4ms
13ms	5ms
14ms	5ms
12ms	5ms
16ms	4ms
20ms	10ms
9ms	3ms
26ms	5ms
12ms	7ms
13ms	5ms
12ms	4ms
15ms	4ms
13ms	5ms
14ms	5ms
12ms	5ms
16ms	4ms
20ms	10ms
19ms	8ms
16ms	6ms
12ms	4ms

15ms	4ms
13ms	5ms
14ms	5ms
12ms	5ms
16ms	4ms
20ms	10ms
9ms	3ms
26ms	5ms
12ms	7ms
13ms	5ms
16ms	4ms
20ms	10ms
19ms	8ms
16ms	6ms
12ms	4ms
15ms	4ms
13ms	5ms
14ms	5ms
12ms	5ms
16ms	4ms
20ms	10ms
9ms	3ms
26ms	5ms
12ms	7ms
13ms	5ms

جدول : 3وقت تنفيذ الخوارزميات

ويكون متوسط زمن تنفيذ الخوارزميات كالتالي:

زمن تنفيذ خوارزمية شجرة القرار: 15.1 ms

زمن تنفيذ خوارزمية العنقدة: 7.6 ms

نلاحظ أن زمالن التنفيذ قليل جداً ولا يزيد عن 16 ميلي ثانية وهذا يعني سرعة كبيرة في استخراج النتائج.

يمكن تلخيص جميع النتائج السابقة بالجدول التالي:

دقة الخوارزمية	الإحكام	الاستدعاء	الحساسية	النوعية	دقة الاختبار	زمن تنفيذ التصنيف	زمن تنفيذ العنقدة
0.814	0.829	0.784	0.89	0.75	0.802	15.1ms	7.6ms

جدول : 4 نتائج عنقدة المنتجات وتصنيف الفئات

6-5-التقنية الثانية:

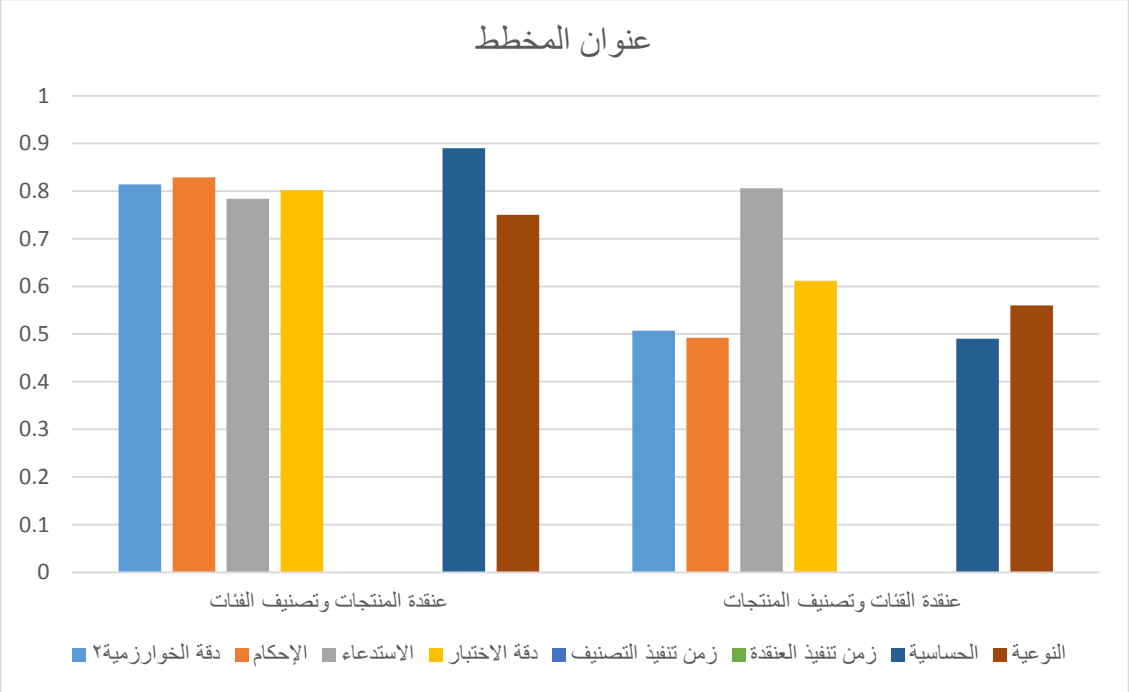
دقة الخوارزمية	الإحكام	الحساسية	النوعية	الاستدعاء	دقة الاختبار	زمن تنفيذ التصنيف	زمن تنفيذ العنقدة
0.507	0.492	0.49	0.56	0.8056	0.612	16.1ms	9.01ms

جدول : 5 نتائج تصنيف المنتجات وعنقدة الفئات

6-6-مقارنة النتائج

بمقارنة النتائج بين الجدولين نجد أن استخدام خوارزمية شجرة القرار في تصنيف الفئات أفضل من استخدامها في تصنيف المنتجات

وذلك بسبب الضجيج الذي قد يصيب خوارزمية شجرة القرار بسبب وجود القيم الرقمية المتباعدة نوعاً ما في البيانات بينما استخدامها في تصنيف المنتجات يلغي هذا الضجيج كون المنتجات جميعها أسماء ولا يوجد فيها قيم رقمية.



شكل : 5 قيم الطريقتين السابقتين

	دقة الخوارزمية	الإحكام	الاستدعاء	دقة الاختبار	زمن تنفيذ التصنيف	زمن تنفيذ العنقدة	الحساسية	النوعية
عنقدة المنتجات وتصنيف الفئات	0.814	0.829	0.784	0.802	15.1ms	7.6ms	0.89	0.75

عنقدة الفئات وتصنيف المنتجات	0.507	0.492	0.8056	0.612	16.1ms	9.01ms	0.49	0.56

جدول : 6 مقارنة الخوارزميات

نلاحظ أن استخدام خوارزمية العنقدة من أجل عنقدة المنتجات حسب مواصفات الزبائن واستخدام خوارزمية شجرة القرار من أجل تصنيف الفئات يعطي نتائج أفضل من استخدام خوارزمية العنقدة من أجل عنقدة الفئات حسب مواصفات الزبائن واستخدام خوارزمية شجرة القرار من أجل تصنيف المنتجات وذلك حسب عدة معايير من معايير قياس دقة خوارزمية التصنيف والتي هي:

1- الدقة
2- الحساسية
3- النوعية
4- الإحكام
5- الاستدعاء
6- دقة الاختبارات
7- السرعة

7-نتائج البحث:

نتيجة للاختبارات التي قمنا بها سابقاً نلاحظ أن استخدام النموذج التنبؤي يعطي أفضل النتائج في توقع فئات المنتجات التي قد يشتري منها الزبون اعتماداً على مواصفات هذا الزبون واعتماداً على المنتجات التي قد اشتراها سابقاً أو المنتجات التي قد اشترها الأشخاص اللذين لديهم نفس المواصفات أو مواصفات قريبة منه وبذلك يمكن التغلب على مشكلة تضخم البيانات بشكل كبير بالنسبة للشركات وعرض مجموعات المنتجات التي تنتمي إلى نفس فئة المنتجات بشكل متقارب والذي سيؤدي بدوره إلى زيادة المبيعات ضمن هذه الفئات.

كما يمكن تصنيف الزبائن حسب مواصفاتهم إلى مجموعات يمكن التنبؤ بفئات المنتجات التي قد يقومون بطلب شراؤها.

8-الآفاق المستقبلية:

نود في تطوير التطبيق بحيث يصبح قادر على التأكد من المنتجات التي سيطلبها المستخدم حسب مواصفاته وتكون نسبة الثقة 100% بحيث تكون الشركات متأكدة تماماً من المنتجات التي يتوقع من المستخدم أن يطلبها حسب موصفاته

9-المراجع

[1] SWATI MAHESH JOSH, "*Market basket analysis using apriori algorithm in data mining*", Department of Information Technology, IRJET, Volume: 05 Issue: 04 Apr-2018.

[2] Riccardo Guidotti_, Giulio Rossetti_x, Luca Pappalardo_x, Fosca Giannotti_, Dino Pedreschix, "*Market Basket Prediction using User-Centric Temporal Annotated Recurring Sequences*", KDD Lab, University of Pisa, Italy, 2016.

[3] X. Su and T. M. Khoshgoftaar, "*A survey of collaborative filtering techniques, Advances in artificial intelligence*", vol. 2009, p. 4, 2009.

[4] C. Chand, A. Thakkar, and A. Ganatra, "*Sequential pattern mining: Survey and current research challenges*" IJSCE, pp. 185–193, 2012.

[5] C.-N. Hsu, H.-H. Chung, and H.-S. Huang,"*Mining skewed and sparse transaction data for personalized shopping recommendation*" ML, vol. 57, no. 1-2, pp. 35–59, 2004.

[6] E. Lazcorreta et al., "*Towards personalized recommendation by two-step modified apriori data mining algorithm*" ESA, pp. 1422–1429, 2008.

[7] P. Wang, J. Guo, Y. Lan, J. Xu, S. Wan, and X. Cheng, "*Learning hierarchical representation model for nextbasket recommendatio,*" in SIGIR. ACM, 2015, pp. 403–412.

[8] Daljeet Kaur, Jagroop Kaur, "*Data Mining in Supermarket: A Survey*", International Journal of Computational Intelligence Research ISSN 0973-1873 Volume 13, Number 8 (2017), pp. 1945-1951.

[9] Zhang Aiguo,Jiang Lanling,Song Ping, "*Application of Data Mining in Supermarket*", International Conference on Consumer Electronics, Communications and Networks (CECNet). IEEE, 2011.

[10] Smita,Priti Sharma," *Use of Data Mining in Various Field:A Survey Paper*", IOSR-Journal of Computer Engineering(IOSR-JCE), 2014.

[11] Deepashri.K.S, Ashwini Kamath, "*Survey on Techniques of Data Mining and its Applications*", International Journal of Emerging Research in Management &Technology ISSN: 2278-9359 (Volume-6, Issue-2), 2017.

[12] Roshan Gangurde, Dr. Binod Kumar, Dr. S. D. Gore, "*Building Prediction Model using Market Basket Analysis"*,International Journal of Innovative Research in Computer and Communication Engineering, Vol. 5, Issue 2, February 2017.