مجلة جامعة طرطوس للبحوث والدراسات العلمية _ سلسلة العلوم الهندسية المجلد (٧) العدد(٣) ٢٠٢٣
Tartous University Journal for Research and Scientific Studies - engineering Sciences Series Vol. (7) No. (3) 2023

تصنيف الزبائن وفق سلوك الشراء باستخدام خوارزمية العنقدة K-means

د. جعفر سلمان*

م. سلمى أديب اليوسف**

(تاريخ الإيداع ٢٠٢٢/١٠/٤ . قُبِل للنشر في ٢٠٢٣/٣/١٥)

□ ملخّص□

تشكل تقنيات التعلم الآلي Machine Learning (ML) الركيزة الأساسية لتطوير التسويق ونظم التجارة الالكترونية، ومع التطور الهائل للتكنولوجيا أصبح لابد من الاستفادة من الكم الهائل من البيانات الواردة يومياً وحتى لحظياً إلى المتاجر، يتيح ذلك التعامل مع العملاء على أنها كنز حقيقي بناء علاقة طويلة الأمد مع العملاء.

يقوم إطار العمل المقترح بتقسيم الزبائن الواردة إلى المتجر، بناءً على سلوك الشراء باعتماد تقنية تعلم الآلة غير الموجه وهو التجميع.

تم تطبيق خوارزمية التجميع K-means على البيانات الناتجة عن تحليل الحداثة والتكرار والقيمة النقدية RFM وذلك عبر معالجة بيانات ضخمة نسبياً لمتجر الكتروني مخصص لبيع الهدايا بالتجزئة عن طريق الانترنت.

وبالنتيجة تم تحسين الشرائح الناتجة عن تحليل RFM عن طريق تطبيق خوارزمية التجميع للوصول إلى تمثيل هادف لقاعدة العملاء قادر على مساعدة المدراء على اتخاذ قرارات تسويقية موجهة بالبيانات.

الكلمات المفتاحية: تجزئة العملاء، العناقيد، تحليل RFM، المسافة الإقليدية، تعلم الآلة.

*مدرس في قسم تكنولوجيا المعلومات . كلية هندسة تكنولوجيا المعلومات والاتصالات . جامعة طرطوس . سوريا

**طالبة ماجستير في قسم تكنولوجيا المعلومات . كلية هندسة تكنولوجيا المعلومات والاتصالات . جامعة طرطوس . سوريا

مجلة جامعة طرطوس للبحوث والدراسات العلمية _ سلسلة العلوم الهندسية المجلد (٧) العدد(٣) ٢٠٢٣
Tartous University Journal for Research and Scientific Studies - engineering Sciences Series Vol. (7) No. (3) 2023

Customer Classification according to purchase behavior using the K-means clustering algorithm

Dr. Jaafar Salman *
Eng. Salma Adib Alyossef **

(Received 4/10/2022 . Accepted 15/3/2023)

□ **ABSTRACT**

Machine Learning (ML) techniques form the mainstay for the development of marketing and e-commerce systems, and with the tremendous development of technology, it is necessary to take advantage of the huge amount of data received daily and even instantaneously to the stores, allows dealing with customers as a real treasure to build a long-term relationship with customers . The proposed framework classifies incoming customers into the store based on their purchase behavior using undirected machine learning technique of aggregation. The K-means aggregation algorithm has been applied to the data generated by RFM analysis by processing relatively large data for an online retail gift shop.

As a result, the segments resulting from RFM analysis were improved by applying the clustering algorithm to reach a meaningful representation of the customer base that is able to help managers make data-driven marketing decisions.

Keywords: Customer Segmentation, Clusters, RFM Analysis, Euclidean distance, Machine Learning

*Teacher, Information Technology Engineering Department, Information and communication Technology Engineering, Tartous University, Syria
**Student Master, Information Technology Engineering Department, Information and communication Technology Engineering, Tartous University, Syria

1- مقدمة:

مع التطور الحالي الهائل الحاصل على مستوى العالم تكنولوجياً، لم تعد أساليب التسويق التقليدية تلبي حاجة الشركات والمؤسسات التجارية، كما لم يعد الوب وسيلة تسويق ثانوية، بل أضحى أساسياً وعلى جميع الشركات ورواد الأعمال تبنيه من أجل إيصال الأفكار والمنتجات بسبب سهولة الوصول من خلاله إلى عدد لامحدود من الزبائن في جميع أنحاء العالم، وبنفس الوقت، مما جعل الوب مصدر رائع للمعرفة بسبب الكم الهائل من البيانات الناتجة عن بيانات العملاء المترددين للمتجر يومياً بل لحظياً.

لابد من معالجة هذه البيانات للاستفادة منها في توجه الشركات وزيادة الإيرادات وذلك عن طريق تنقيب الويب والذي يسهم في تحويل هذه البيانات الخام إلى معرفة تساهم في اتخاذ القرارات، والتي أصبحت أدوات أساسية في الشركات الرائدة على مستوى العالم مثل Google وأمازون Amazon و لتحسين استراتيجيات التسويق والمنتجات لتصبح على هذا المستوى من المنافسة.

الحلول التي تم تحقيقها في هذا البحث هي عمليات تعدين جديدة للبيانات، وموجهة بتنقيب الوب، والسماح في الوقت نفسه بتغذية عكسية وفقًا للمعرفة المستمدة من عملية التعدين.

2-الدراسات المرجعية:

تناولت الدراسات السابقة نظرة عامة على أنواع تنقيب بيانات الويب وبعض تطبيقاتها بالإضافة إلى المراحل الأساسية في عملية التنقيب من جمع البيانات وصولاً لصياغة المعرفة كما طرح بعض تحديات تنقيب الويب في مجال التجارة الالكترونية[1].

كما قُدِمَ مؤخراً دراسة تم من خلالها استخدام تحليل RFM لتجزئة المنتجات كما تم مقارنة الأساليب الإحصائية لثماني تقنيات مختلفة من أجل تحديد العدد الأمثلي للعناقيد دون النظر إلى الأساليب النوعية المعتمدة على خبرة المجال والتي من شأنها أن تحسن النتائج.[2]

استخدم Christodoulakis في دراسة بعنوان Customer Clustering using RFM analysis نموذج Pyramid لتجميع العملاء من خلال الإيرادات التي حققوها لتحسين علاقات العملاء مع عملاء البنوك. [3]

وفي عام ٢٠١٥ قام مجموعة من الباحثين بإجراء دراسة لتحديات البيانات الكبيرة وخوارزميات العنقدة وأنواعها بعنوان Big data clustering: Algorithms and challenges حيث توصلت إلى أن خوارزميات التنقيب عن البيانات الكلاسيكية تحتاج إلى تحسين ودعم من خلال المعالجة المسبقة للبيانات من أجل تقليل حجمها. [٤]

أنجز الباحثان Valarmathy.N, Krishnaveni.S دراسة بعنوان Performance Evaluation and Comparison of Clustering Algorithms used in Educational Data Mining تناولت تطبيق خوارزميات التجميع في الأنظمة التعليمية التقليدية. [٥]

3- هدف البحث وأهميته:

يهدف البحث إلى تقديم إطار عمل قادر على توجيه منظومة الأعمال نحو بناء استراتيجية تقوم على تفضيلات الزبائن وتوجهاتهم، وذلك من خلال دراسة وفهم نمط الشراء والسلوك الذي يظهره العملاء للمتجر الالكتروني من خلال بيانات حقيقية التي تم جمعها خلال فترة زمنية معينة، وذلك بتطبيق خوارزمية العنقدة Kmeans على البيانات الناتجة عن تحليل RFM واقتراح التوصيات والإجراءات التي يمكن للشركة إجراءها على أساس الشرائح الناتجة.

وتأتي أهمية البحث من خلال مايلي:

- تطبيق المعرفة البشرية وتقنيات الذكاء الاصطناعي لإدارة ودعم القرار في عالم الأعمال.
- الاستفادة من المعلومة حيث تقود للكفاءة أي تحقيق السرعة والدقة والموضوعية في إدارة الأعمال التنظيمية والفردية مما يقود إلى الحلول الإبداعية في الشركات.

4- منهجية البحث:

اقتراح بنية جديدة لتوسيع متجر الكتروني بحيث يكون منصة بيع وتسويق ذكية تتناسب مع حاجات وتفضيلات الزبائن عن طريق تطبيق تقنيات تنقيب الوب.

وتم دراسة ومعالجة عينات الزبائن تم إجراء الدراسة على بيانات الزبائن التي تم الحصول عليها من موقع Kaggle[6] حيث تحتوي على ٥٤١٩٠٩ عينة باستخدام برنامج Anaconda و لغة البرمجة Python3 وتم تصميم إطار العمل بناء على خطوات واضحة، وهي معالجة البيانات وتطبيق تقسيم RFM ثم خوارزمية العنقدة K-means من أجل تقسيم الزبائن وفق سلوك الشراء، بحيث يمكن لمدير التسويق التعامل مع الزبائن وفق سمات الشراء لديهم واتخاذ الإجراءات التسويقية المتناسبة مع توجهات الزبائن، وذلك وفق الخطوات التالية:

4-1 جمع البيانات:

تم العمل على قاعدة بيانات لمتجر تجزئة الكتروني لبيع الهدايا في المملكة المتحدة قدمها الدكتور Daqing Chen [٧] في UCI Machine Learning Repository ، وموجودة على موقع kaggle [6].

تحوي قاعدة البيانات على جميع المعاملات التجارية الفعلية التي حدثت في عام ٢٠١١ وتحديداً بين ٢٠١٠/١٢/٠١ و ٢٠١١/١٢/٠٩.

كما يوضح الجدول التالي الميزات الموجودة في قاعدة البيانات المستخدمة لأغراض بحثنا.

الجدول (١) جزء من قاعدة بيانات المتجر الالكتروني المقدم من UCI Machine Learning Repository

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitePrice	CustomerID	Country
٠	٥٣٦٣٦٥	A٨٥١٢٣	WHITE HANGING HEART T-LIGHT HOLDER	٦	12/10/2010 8:26	٢.٥٥	١٧٨٥٠.٠	United Kingdom
١	٥٣٦٣٦٥	٧١٠٥٣	WHITE METAL LANTERN	٦	12/10/2010 8:26	٣.٣٩	١٧٨٥٠.٠	United Kingdom
٢	٥٣٦٣٦٥	B٨٤٤٠٦	CREAM CUPID HEARTS COAT HANGER	٦	12/10/2010 8:26	٢.٧٥	١٧٨٥٠.٠	United Kingdom
٣	٥٣٦٣٦٥	G٨٤٠٢٩	KNITED UNION FLAG HOT WATER BOTTLE	٦	12/10/2010 8:26	٣.٣٩	١٧٨٥٠.٠	United Kingdom
٤	٥٣٦٣٦٥	29E٨٤٠	RED WOOLY HOTTLE WHITE HEART	٦	12/10/2010 8:26	٣.٣٩	١٧٨٥٠.٠	United Kingdom

4-2 تنظيف البيانات:

يجب أن تكون البيانات المشكلة لدخل عملية التنقيب نظيفة حتى لا يتأثر الخرج بالبيانات غير المناسبة لذا تم ما يلي:

١- حذف السجلات التي تحوي قيم فارغة.

٢- تؤدي الكمية الكبيرة من البيانات الزائدة إلى إبطاء عملية اكتشاف المعرفة أو إرباكها لذا تم حذف السجلات المكررة.

٣- تؤثر القيم السالبة بشكل مباشر في النتائج لذا تم حذف الصفوف التي تحوي على قيم سالبة في عمود الكمية Quantity والسعر UnitPrice.

يظهر الشكل معلومات قاعدة البيانات بعد مرحلة تنظيف البيانات يظهر فيه انخفاض عدد سجلات البيانات من ٥٤١٩٠٨ إلى ٣٩٧٨٨٤ ، كما يظهر عدم وجود قيم فارغة.

```
Int64Index: 397884 entries, 0 to 541908
Data columns (total 8 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   InvoiceNo    397884 non-null  object
 1   StockCode    397884 non-null  object
 2   Description  397884 non-null  object
 3   Quantity     397884 non-null  int64
 4   InvoiceDate  397884 non-null  datetime64[ns]
 5   UnitPrice    397884 non-null  float64
 6   CustomerID   397884 non-null  float64
 7   Country      397884 non-null  object
dtypes: datetime64[ns](1), float64(2), int64(1), object(4)
```

الشكل(١) معلومات البيانات بعد عملية تنظيف البيانات

4-3 تقسيم الزبائن وفق سلوك الشراء:

تعتبر معرفة العملاء خطوة أساسية من أجل تبني الاستراتيجيات الصحيحة التي تقوم بتوفير المال وتوجيه الجهود ومن طرق التعرف على العملاء هو تقسيم السوق إلى مجموعات عملاء منفصلة بحيث تتشابه خصائص العملاء ضمن نفس المجموعة، ويعتبر وسيلة قوية لتحديد احتياجات العملاء، مما يمكن الشركات من التفوق على المنافسة من خلال تطوير منتجات وخدمات متناسبة بطريقة مدروسة تتلاءم مع هذه الاحتياجات.

تمت عملية تقسيم الزبائن إلى شرائح ذات معنى من خلال مرحلتين:

٤-٣-١ المرحلة الأولى تحليل الحداثة والتكرار والقيمة النقدية:

تم اقتراح تحليل RFM لأول مرة بواسطة A.M. Huges لتجزئة العملاء وتحليل السلوك في ١٩٩٤ [٨]، تحليل RFM هو أسلوب تسويقي فعال يستخدم لتصنيف العملاء وتجميعهم كميًا بناءً على حداثة معاملاتَهم التجارية وتكرارها وإجمالي قيمتها النقدية [٩]، كما يستخدم لتحديد أفضل العملاء وتنفيذ حملات تسويقية مستهدفة.

يقوم النظام بتعيين درجات رقمية لكل عميل بناءً على هذه العوامل لتقديم تحليل موضوعي.

تحديد تعريفات قيم RFM بالنسبة للبيانات المقترحة (هندسة الميزات):

١- الحداثة R: عدد الأيام منذ آخر معاملة للعميل، كلما صغر كان أفضل للشركة لأن العميل يعتبر نشيط (تاريخ اليوم الذي يلي آخر معاملة متوفرة في قاعدة البيانات - تاريخ آخر معاملة للعميل)

٢- التكرار F: عدد المعاملات (عدد الفواتير) التي قام بها العميل في آخر ١٢ شهر.

٣- القيمة النقدية M: القيمة الإجمالية التي صرفها العميل في آخر ١٢ شهر(سعر المنتج *الكمية).

١٢ شهر هي قيمة معيارية يتم اختيارها بناء على (نموذج العمل، دورة حياة المنتج، العملاء) تم تحديدها في بحثنا بناء على الداتا الموجودة، يوضح الجدول(٢) لقطة من قاعدة البيانات بعد استنتاج قيم الحداثة والتكرار والقيمة النقدية، حيث يعرض قيمة كل من الميزات الثلاثة بالنسبة لكل عميل على حدا.

الجدول (٢) جزء من قاعدة البيانات يوضح الوصول لقيم الحداثة والتكرار والقيمة النقدية لكل عميل

CustomerID	Recency	Frequency	Monetary Value
١٢٣٤٦.٠	٣٢٦	٢	٠.٠٠
١٢٣٤٧.٠	٢	١٨٢	٤٣١٠.٠٠
١٢٣٤٨.٠	٧٥	٣١	١٧٩٧.٢٤
١٢٣٤٩.٠	١٩	٧٣	١٧٥٧.٥٥
١٢٣٥٠.٠	٣١٠	١٧	٣٣٤.٤٠

بناء شرائح RFM:

يقسم RFM العملاء بناءً على المقاييس الثلاثة الرئيسية بحيث تم إعطاء درجة من1 إلى 4 حسب ما يحققه العميل بالنسبة للمقياس، وقد تستخدم التطبيقات المختلفة لنظام تحليل RFM قيماً ومقاييس مختلفة.

تم تقسيم العملاء بناءً على النسبة المئوية بحيث يأخذ العميل بالنسبة للتكرار القيمة 4 إذا حقق تكراراً يقع في أعلى 25% من قيم المقياس و القيمة 3 إذا حقق تكرار ينتمي لل25% التالية و1 لأدنى 25%، وهكذا بالنسبة لبقية المقاييس.

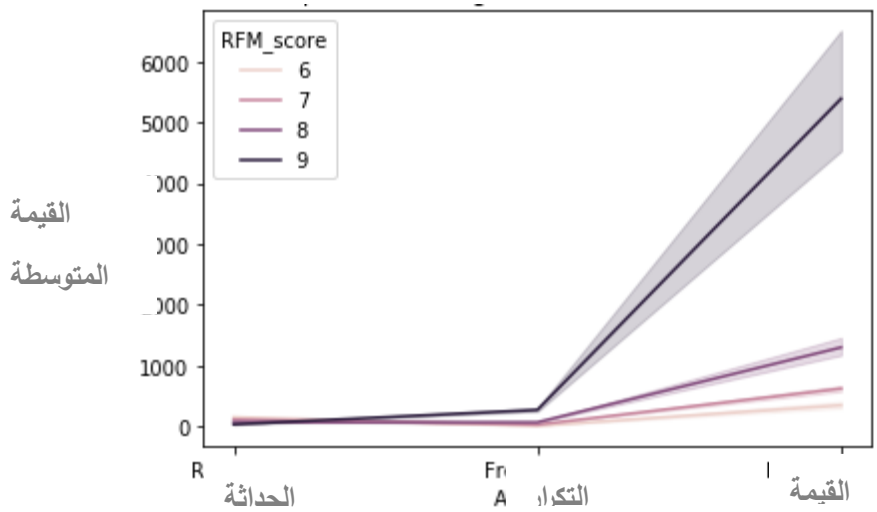
يتم دمج قيم R و F و M من أجل تعيين تصنيف الزبون حيث يُطلق على مجموعة القيم الثلاث لكل عميل اسم خلية RFM، تقوم المؤسسات بتوسيط هذه القيم معًا، ثم فرز العملاء من الأعلى إلى الأدنى للعثور على العملاء الأكثر قيمة ويمكن أن تقوم الشركات بوزن القيم بشكل مختلف.

مثلاً إذا حقق الزبون 513 هذا يعني: [10]

الحداثة R=5: أجرى الزبون عملية شراء مؤخراً، التكرار F= ١: يشتري الزبون منتجات من الشركة بمعدل تكرار منخفض جداً، القيمة النقدية M=3: يعتبر إنفاق الزبون متوسط.

وتم تلخيص القيم الثلاثة في RFM_Score والتي تعتبر قيمة نسبية للعميل بالنسبة للعوامل الثلاثة حيث تم الحصول عليها عن طريق جمع ماحققه بالنسبة للمقاييس الثلاثة فإن RFM_Score بالنسبة للعميل السابق هو 5+1+3 أي 9 .

يوضح الشكل (2) الشرائح الناتجة عن تطبيق تقسيم RFM على قاعدة البيانات السابقة حسب RFM_score، وهي حيث تم تقسيم العملاء إلى ٤ شرائح حققت القيم النسبية ٦، ٧، ٨ ، ٩:



الشكل(2)الشرائح الناتجة عن تقسيم RFM حسب RFM_score

يوضح الشكل أن خرج عملية التقسيم وفق تحليل هو شرائح غير واضحة الملامح، التمايز بين الشرائح على أساس القيمة النقدية فقط.

تم استخدام العنقدة لتحسين التجزئة في تحليل RFM ليكون أكثر موضوعية ودقة باستخدام خوارزمية K-means بحيث تحقق مجموعات العملاء التشابه الأمثلي للعناصر ضمن عملاء المجموعة الواحدة.

٤-٣-٢ تطبيق خوارزمية التجميع على قيم RFM:

التجميع هو أسلوب تعلم آلي غير موجه لا يحتاج معلومات مسبقة عن المجموعات أو عن خصائصها، يتضمن تقسيم مجموعة من البيانات إلى مجموعات بناء على التشابه بين العناصر بحيث تكون المسافة بين نقاط البيانات في مجموعة منخفضة جدًا مقارنة بالمسافة بين مجموعتين، بمعنى آخر يتشابه أعضاء المجموعة الواحدة، بينما تختلف العناصر التي تنتمي إلى مجموعات منفصلة عن بعضها .

١- اختيار خوارزمية العنقدة:

هناك خوارزميات مختلفة للتجميع بناء على طريقتها في بناء العنقود [١2]، بما أن البيانات كبيرة الحجم يجب أخذ السرعة بعين الاعتبار، كما يجب اختيار خوارزمية تأخذ جميع نقاط البيانات بعين الاعتبار ومراعاة البساطة ليتمكن أصحاب المصلحة من فهم العمل والنتائج، لذا تم استبعاد خوارزميات التجميع التي تتسم بالبطء في العمل، كما الخوارزميات التي تتعامل مع المجموعات غير الكثيفة على أنها ضوضاء [13]، وتم اختيار خوارزمية Kmeans وهي خوارزمية سهلة التنفيذ نسبيًا، كما أنها سريعة بالنسبة لغيرها من خوارزميات التجميع خاصة عند العمل مع البيانات الكبيرة، بالإضافة إلى سهولة التكيف مع الأمثلة الجديدة واتخاذها بالاعتبار العناقيد بمختلف أشكالها وأحجامها.

خوارزمية العنقدة Kmeans:

من أشهر خوارزميات العنقدة المعتمدة على التقسيم، وهي خوارزمية تكرارية يتم فيها تحديد عدد العناقيد K مسبقاً، تقوم بتجميع البيانات ضمن عناقيد بحيث تكون البيانات ضمن العنقود الواحد ذات خصائص مشتركة لكن تختلف عن البيانات في العناقيد الأخرى[14].

وتقوم الخوارزمية بتعيين كل نقطة بيانات بشكل متكرر إلى إحدى مجموعات K بناءً على تشابه الميزات[3] أي تقليل مجموع مربعات المسافات بين النقطه الوسطى للعنقود ونقاط البيانات المرتبطة بها.

تتضمن خطوات خوارزمية العنقدة Kmeans مايلي:

الخطوة ١: تحديد عدد العناقيد k.

الخطوة ٢: تحديد k من النقاط العشوائية من البيانات كنقاط مركزية centroids أي تعيين مراكز العناقيد الأولية.

الخطوة ٣: تعيين كل نقطة من نقاط البيانات لأقرب مركز عنقود عن طريق حساب المسافة الإقليدية التابع المستخدم من أجل قياس التشابه هو تابع المسافة الإقليدية[14]:

$$Minimize \sum_{j=1}^{k} \sum_{i=1}^{n} (x_{ij}^{(1)} - c_j)^2$$

k: عدد العناقيد، n: عدد نقاط البيانات، c: مركز العنقود (النقطة الوسطى)، x_{ij}: نقطة البيانات.

الخطوة ٤: إعادة حساب النقط الوسطى للعناقيد وهي متوسط العناصر ضمن كل عنقود.

الخطوة ٥: تكرار الخطوتين ٣ و ٤ حتى التوقف عند تحقيق أحد المعايير:

١- ثبات النقط الوسطى دون تغيير. ٢- بقاء النقاط في نفس المجموعة. ٣- الوصول إلى الحد الأقصى لعدد التكرارات.

٢- المعالجة المسبقة للبيانات:

خوارزمية K–Means حساسة بالنسبة للقيم المتطرفة والقيم الشاذة لأنه إذا انتمى عنصر شاذ إلى أحد العناقيد سوف يحرف المتوسط بشكل كبير وبالتالي سوف يؤثر على عملية توزيع العناصر ضمن العناقيد.

تعد المعالجة المسبقة للبيانات مرحلة حرجة، حيث تتضمن أن تلبي البيانات جميع متطلبات الخوارزمية لتنفيذ وتقديم نتائج ذات مغزى، تفترض خوارزمية Kmeans مجموعة من الافتراضات والشروط في البيانات من أجل العمل بالشكل الأمثلي وتقارب أفضل.

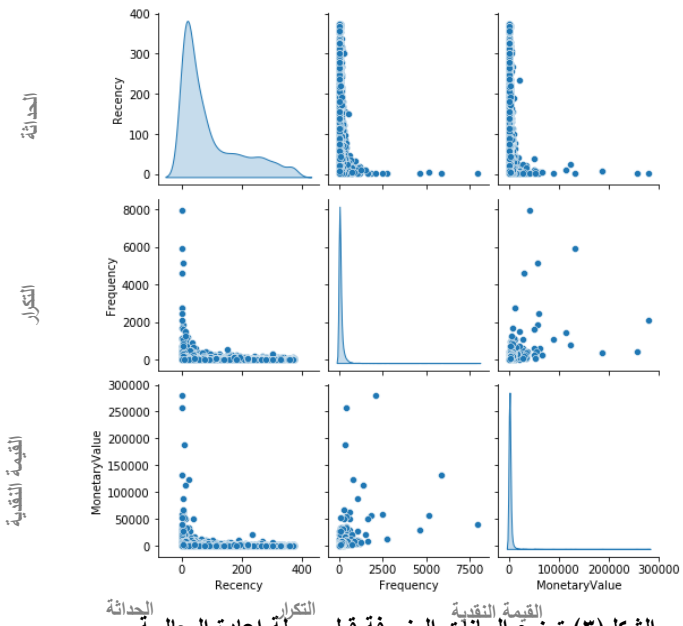
١) جميع المتغيرات لها توزيعات طبيعية:

تم إزالة انحراف التوزيع عن طريق تطبيق تحويل boxcox والذي يعطى من أجل القيم الموجبة وفق المعادلة التالية:

$$y(\lambda) = \begin{cases} \dfrac{y^{\lambda} - 1}{\lambda}, & \text{if } \lambda \neq 0 \\ \log y, & \text{if } \lambda = 0 \end{cases} \qquad (٢)$$

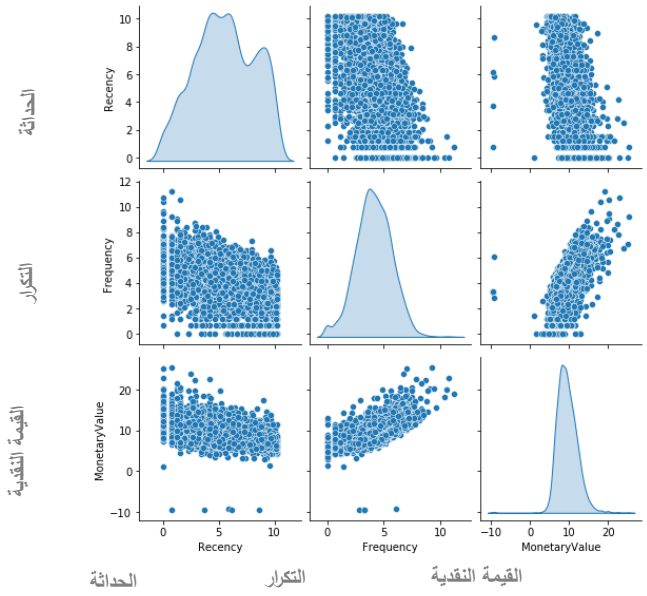
يتم أخذ جميع قيم λ في الاعتبار ويتم تحديد القيمة المثلى للبيانات وهي القيمة التي ينتج عنها أفضل تقريب لمنحنى التوزيع الطبيعي.

يوضح الشكل (٣) توزيع قيم الميزات التي تم الوصول إليها عبر تحليل الحداثة والتكرار والقيمة النقدية قبل تطبيق تحويل boxcox حيث نلاحظ أن توزيع المتغيرات منحرف إلى اليسار.



الشكل(٣) توزيع البيانات المنحرفة قبل مرحلة إعادة المعالجة

كما يلاحظ من خلال الشكل (٤) والذي يعبر عن توزيع البيانات بعد تطبيق تحويل boxcox حيث يظهر لدينا منحني غير متناظر بشكل مثالي لكن فيه انحراف بسيط جداً مقارنة بالتوزع الأصلي.



الشكل(٤) تطبيق Box_cox لإزالة تشوه البيانات

٢) تحجيم الميزاتFeature Scaling:

تختلف مجالات قيم المتغيرات العددية للميزات لذا لابد من تحجيم الميزات من أجل بناء نموذج تعلم آلي فعال ومساعدة خوارزمية التعلم الآلي على فهم العلاقة النسبية بين المتغيرات وإلا ظهرت نتائج غير مفهومة.

تم استخدام طريقة التحجيم القياسي Standard Scaler :

نعني بالتحجيم القياسي نقل جميع البيانات إلى توزيع عادي قياسي بمتوسط صفري وانحراف معياري يساوي الواحد حيث يتم حساب القيم وفق المعادلة:

$$Z = \frac{x_i - \mu}{\sigma} \qquad (3)$$

حيث: Z هو المتغير المراد تحجيمه، μ هو المتوسط الحسابي لقيم z، الانحراف المعياري للمتغير.

تم توحيد المتوسط والانحراف المعياري أي تنفيذ عمليات تمركز data Centering وتحجيم الميزات feature Scalingعن طريق تابع StandardSaler من مكتبة scikit_learn في بايثون والذي يعيد مصفوفة بدلاً من إطار بيانات (kmeans تعمل بشكل أفضل مع نوع البيانات ndarray بدلاً من dataframes).

يوضح الجدول(٣) القيم الإحصائية للبيانات حيث يلاحظ القيم الموحدة للمتوسط الحسابي والانحراف المعياري للميزات :

الجدول (٣) القيم الإحصائية للبيانات المعاد معالجتها

	Recency	Frequency	Monetary Value
count	٤٣٢٢.٠٠	٤٣٢٢.٠٠	٤٣٢٢.٠٠
mean	٠.٠٠	٠.٠٠	٠.٠٠
std	1.00	1.00	1.00
min	−0.07	−2.63	−7.27
25%	−0.70	−0.65	−0.67
50%	٠.٠١	−0.01	−0.11
75%	٠.٨٣	٠.٦٩	٠.٥٩
max	١.٧٧	٤.٥٢	٦.١٠

٣- اختيار عدد العناقيد:

هناك العديد من الطرق المستخدمة لتحديد العدد الصحيح والأمثل للعناقيد[2]، لكن لا يوجد طريقة أفضل من غيرها ولكن بديل عنها.

استخدام طريقة الكوع Elbow Method[3]: وهي طريقة بصرية تنظر إلى التباين الكلي داخل العنقود أو إجمالي مجموع مربعات الأخطاء داخل العناقيد SSE كتابع لعدد العناقيد حيق كلما زاد k صغر مجموع SSE والعكس.

الفكرة الأساسية لطريقة الكوع هي أنه كلما زاد عدد العناقيد k سيصبح تقسيم العينة أكثر دقة، وبالتالي سيصبح مربع الخطأ أصغر تدريجيًا بشكل طبيعي.

وعندما يصل k إلى عدد المجموعات الأمثلي، فإن القيمة التي يتم الحصول عليها عن طريق زيادة k تسبب انخفاض حاد في قيمة مربع الخطأ، وبالتالي فإن SSE Sum of squared error و سينخفض بشكل حاد، ثم يتسطح تدريجيًا مع استمرار زيادة قيم k ، مما يعني أن العلاقة بين SSE وk على شكل كوع، وقيمة k عند هذا الكوعي هي عدد العناقيد الأمثلي في البيانات لهذا السبب تسمى الطريقة طريقة الكوع.

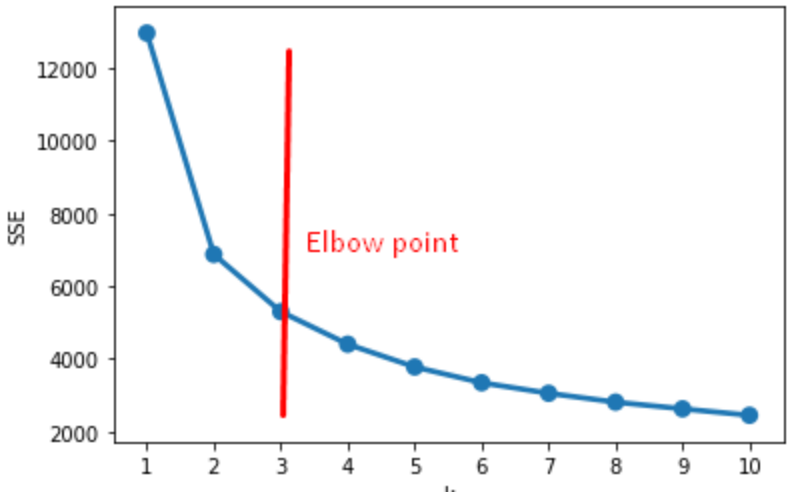
تم اختيار مجال من القيم لعدد العناقيدk (من ١ إلى ١١) ومن أجل كل قيمة ل k تقوم بالتنفيذ على كامل مجموعة البيانات وحساب sum of squared Error أي مجموع مربعات الأخطاء.

من أجل كل عنصر يكون الخطأ E هو المسافة إلى مركز العنقود

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} (p - m_i)^2 \quad (4)$$

حيث K هو عدد العناقيد الناتجة، C مجموعة العناقيد الناتجة {C=ci,c2,.....,c1، P مجموعة العناصر ضمن العنقود، m هو مركز العنقود.

يوضح الشكل (3) الرسم البياني الناتج عن تطبيق طريقة الكوع بحيث يوضح الزاوية الأكبر التي يتم بعدها تناقصSSE التدريجي مع زيادة عدد العناقيد.



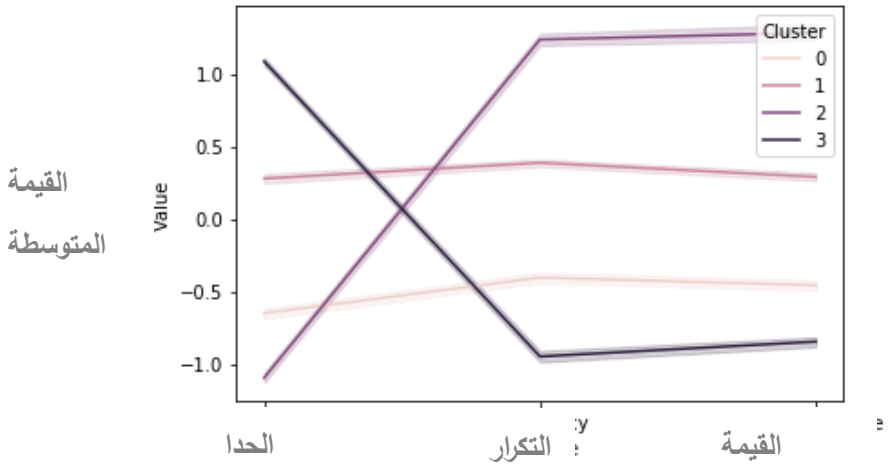
الشكل (٣) تطبيق طريقة الكوع لحساب عدد العناقيد

الخيار الأفضل هو نقطة الكوع أو النقطة التي بعدها لذلك سنختار K=4.

٤- تطبيق خوارزمية K-mean:

تم تقسيم مجموعة معينة من العملاء إلى ٤ مجموعات متمايزة، وتعتبر هذه المجموعات هي الأقسام الأربعة من العملاء التي تم اكتشافها وفق سلوك الشراء.

يوضح الشكل (٥) العناقيد الأربعة الناتجة عن تطبيق خوارزمية العنقدة K-means مع قيم الحداثة والتكرار والقيمة النقدية المميزة لها.



الشكل(٥) تطبيق خوارزمية K-means

٣- تفسير المجموعات واقتراح التوصيات:

من أجل اكتشاف سمات الشرائح قمنا بحساب القيم المتوسطة للميزات التي تم بناء الشرائح بناء عليها، والتي ترتبط بسلوك المستخدم الشرائي وهي الحداثة والتكرار والقيمة النقدية كما في الصورة السابقة.

والناتج هو ٤ شرائح وهي:

الشريحة الأولى العنقود ٠ الزبائن الذين على وشك المغادرة :

مؤشرات الأداء الثلاثة منخفضة وبالتالي يجب وضع استراتيجيات مخصصة، فيما يلي بعد التوصيات التي يمكن للشركة اتخاذها من أجل اكتسابهم وتحويلهم عملاء محتملين للشركة.

١- تواصل من أجل معرفة معوقات الشراء.

٢- رسائل مخصصة تدعوهم لاتخاذ إجراء الشراء.

٣- مكافآت في حال شراء عدة منتجات من أجل تشجيعهم.

الشريحة الثانية العنقود ١ العملاء المخلصين:

بسبب القيم العالية التي يحققونها بالنسبة لمؤشرات الأداء الرئيسية الثلاثة فهم بحاجة عناية خاصة وتقدسم حوافز لزيادة الإنفاق أي القيمة النقدية الخاصة بهم من خلال:

١- الحسومات.

٢- عروض حصرية على المنتجات.

٣- استهدافهم في الحملات الإعلانية.

هذا التحفيز قد يحولهم إلى زبائن مخلصين وسيزيد من القيمة النقدية والتكرار مما سيسهم في زيادة إيرادات الشركة.

الشريحة الثالثة العنقود٢ فئة العملاء الأفضل للأعمال:

وهم العملاء الذين يحققون تردد وقيمة نقدية عالية، لذا فهم المساهمون الأكبر في إيرادات الشركة، من بين التوصيات الأساسية التي على الشركة اعتمادها هي الاستفادة من هذه الشريحة في الترويج للعلامة التجارية للشركة عبر:

١- أخذ شهاداتهم من أجل الدعاية.

٢- مكافأتهم بتقديم خدمات في حال الحصول على زبائن عن طريقهم.

٣-عروض خاصة من أجل زيادة دافع الشراء وشعورهم بالتقدير وبناء الولاء.

٤-استطلاع من أجل تعزيز ودعم المنتجات التي ساهمت في اكتسابهم.

٥- العمل على منتجات جديدة.

٦- اقتراح منتجات مرتبطة بالمنتجات السابقة.

تجدر الإشارة هنا إلى أن التحفيز يمكن أن يسهم في نفورهم وبالتالي انخفاض الإيرادات.

الشريحة الرابعة العنقود٣ الزبائن الجدد:

وهم العملاء الذين يحققون قيمة حداثة عالية لذا على الشركة اكتسابهم من خلال المتابعة المستهدفة التي سوف تحولهم إلى عملاء مخلصين، لذا تقدم لهم حوافز لاتخاذ قرار الشراء مرة أخرى كالرسائل الترحيبية والهدايا.

٥- النتائج:

من خلال العمل على بيانات حقيقية لمتجر الكتروني مكونة من ٥٤١٩٠٩ سجل تجاري تم التركيز على الأنماط السلوكية للزبائن والحصول على شرائح سهلة التفسير ذات جودة عالية من خلال تطبيق تحليل RFM لاستخلاص السمات المرتبطة بسلوك الشراء للمستخدم وهي نشاطه (الحداثة) ومعدل تردده إلى المتجر (التكرار) ومعدل إنفاقه (القيمة النقدية).

وبالنتيجة فإن تطبيق خوارزمية التجميع k-means على تقسيم RFM جعل شرائح العملاء أكثر فائدة في الأعمال، وحقق الهدف الأساسي للتجزئة وهو تمثيل هادف لقاعدة العملاء في مجموعات متشابهة يمكن تفسيرها واستخدامها في تخصيص التسويق أو طرح المنتجات من خلال التركيز على الأنماط السلوكية الخاصة بكل مجموعة وصياغة رسائل محددة متناسبة مع هذه الأنماط السلوكية للعملاء.

٦- المراجع:

[1] DESHPANDE. S. (2012) A Survey on Web Data Mining Applications. *IJCA Proceedings on Emerging Trends in Computer Science and Information Technology*

[٢] GUSTRIANSYAH, R; SUHANDI, N; Antony, F,2020. *Clustering optimization in RFM analysis based on k-means*. *Indones. J. Electr. Eng. Comput. Sci*, *18*(1), 470-477.

[٣]SYAKUR, M. A; KHOTIMAH, B. K; ROCHMAN, E. M. S; SATOTO, B. D. (2018, April), *Integration k-means clustering method and elbow method for identification of the best customer profile cluster*. *IOP conference series: materials science and engineering* (Vol. 336, No. 1, p. 012017). IOP Publishing.

[4] ZERHARI, B., LAHCEN, A. A., & MOULINE, S. (2015, May). *Big data clustering: Algorithms and challenges.* IOP. *Conf. on Big Data, Cloud and Applications* (BDCA'15).

[5] VALARMATHY, N., KRISHNAVENI, S. (2019). *Performance evaluation and comparison of clustering algorithms used in educational data mining*. *International Journal of Recent Technology and Engineering*, ISSN, 2277-3878.

[6] https://www.kaggle.com/

[٧] Dr. Daqing Chen, Director: *Public Analytics group*. chend '@' lsbu.ac.uk, School of Engineering, London South Bank University, London SE1 0AA, UK.

[٨] HUGHES, A, M.١٩٩٤, *Strategic Database Marketing*. Probus Publishing Company.

[٩] CHRISTY,A, J; UMAMAKESWARI, A; PRIYATHARSINI, L; NEYAA, A, 2021, *RFM ranking–An effective approach to customer segmentation*. *Journal of King Saud University-Computer and Information Sciences*, 33(10), 1251-1257.

[١٠] SAHA, S., & BANDYOPADHYAY, S. (2012). *Some connectivity based cluster validity indices. Applied Soft Computing*, 12(5), 1555-1565.

[12]COMIN, C, H; CASANOVA, D; BRUNO, O, M; AMANCIO, D, R; COSTA, L, F; RODRIGUES, F, A,2019, *Clustering algorithms: A comparative approach*. *PloS one*, VOL.*14*,NO.1,122)

[13] ESTER, M; KRIEGEL, P; SANDER, J; XU, X, (1996, August), *A density-based algorithm for discovering clusters in large spatial databases with noise*. In *kdd*.Vol. 96, No. 34, pp. 226-231

[14] SINATRYA, N. S., & WARDHANI, L. K. (2018, August). *Analysis of K-Means and K-Medoids's Performance Using Big Data Technology.* In 6th International Conference on Cyber and IT Service Management (CITSM) (pp. 1-5). IEEE.