E-ISSN: | P-ISSN:

Management Sciences and Decision Analysis

www.msda.reapress.com

Manag. Sci. Decis. Anal. Vol. 1, No. 2 (2023) 92–101.

Paper Type: Original Article

Presenting a Method for Improving Online Store Sales Based on Customer Behavior Analysis Using Data Mining

Fataneh Tabarteh Farahani*

Department of Computer Science, Faculty of Engineering, Arak Branch, Islamic Azad University, Arak, Iran; tfarahani.f@gmail.com.

Citation:

Tabarteh Farahani, F. (2023). Presenting a method for improving online store sales based on customer behavior analysis using data mining. *Management sciences and decision analysis, 1*(2), 92-101.

Received: 01/06/2023 Reviewed: 19/09/2023 Revised: 09/10/2023 Accepted: 08/11/2023

Abstract

Purpose: Given the rapid growth of e-commerce and the importance of customer behavior analysis in improving sales, this study aims to present a data mining model based on fuzzy clustering. The necessity of this research arises from the need of organizations to identify behavioral patterns and enhance the effectiveness of sales strategies.

Methodology: This study employs the CRISP data mining methodology. The process includes data preprocessing, normalization, clustering using the K-Means algorithm, and feature selection via the Fuzzy C-Means algorithm. The dataset is extracted from the UCI repository and contains behavioral attributes of online shoppers.

Findings: The results showed that the proposed model achieved 96.70% accuracy, outperforming other algorithms. Evaluation metrics included accuracy, classification error, precision, recall, and F-measure. Influential features in customer purchasing behavior were identified and categorized.

Originality/Value: The innovation of this study lies in combining K-Means and FCM algorithms within the CRISP framework to analyze customer behavior. The model is generalizable to other datasets and can be used to optimize marketing and sales decisions.

Keywords: Customer behavior, Data mining, Fuzzy clustering, K-Means algorithm, Online store.

Corresponding Author: tfarahani.f@gmail.com https://doi.org/10.22105/msda.v1i2.125

علوم مدیریت و تحلیل تصمیم

دوره ۱، شماره (۲)، (۱۴۰۲)، ۱۰۱-۹۲

www.msda.reapress.com

نوع مقاله: پژوهشی

ارایه یک روش بهبود میزان فروش فروشگاه های آنلاین بر اساس تحلیل رفتار مشتری با استفاده از داده کاوی

فتانه تبرته فراهانی*

گروه علوم کامپیوتر، دانشکده فنی و مهندسی، واحد اراک، دانشگاه آزاد اسلامی، اراک، ایران.

چکیده

هدف: با توجه به رشد روزافزون تجارت الکترونیک و اهمیت تحلیل رفتار مشتریان در بهبود فروش، این مطالعه با هدف ارایه یک مدل داده‌کاوی مبتنی بر خوشه‌بندی فازی انجام شده است. لزوم انجام این تحقیق از نیاز سازمان‌ها به شناسایی الگوهای رفتاری مشتریان و ارتقا اثربخشی استراتژی‌های فروش ناشی می‌شود.

روش‌شناسی پژوهش: در این پژوهش از متدولوژی *CRISP* برای داده‌کاوی استفاده شده است .مراحل شامل پیش‌پردازش داده‌ها، نرمال‌سازی، خوشه‌بندی با الگوریتم *K-Means* و انتخاب ویژگی‌های مؤثر با الگوریتم *Fuzzy C-Means* بوده‌اند. داده‌ها از پایگاه *UCI* استخراج شده و شامل ویژگی‌های رفتاری کاربران در فروشگاه‌های آنلاین هستند.

یافته‌ها: نتایج نشان داد که مدل پیشنهادی با دقت ۹۶/۷۰% عملکرد بهتری نسبت به سایر الگوریتم‌ها دارد .معیارهای ارزیابی شامل دقت، خطای طبقه‌بندی، *Precision*، *Recall* و F-measure بوده‌اند. همچنین ویژگی‌های موثر بر رفتار خرید مشتریان شناسایی و طبقه‌بندی شدند.

اصالت/ارزش افزوده علمی: نوآوری این پژوهش در ترکیب الگوریتم‌های *K-Means* و *FCM* در چارچوب *CRISP* برای تحلیل رفتار مشتریان است. این مدل قابلیت تعمیم به سایر مجموعه داده‌ها را داشته و می‌تواند در بهینه‌سازی تصمیمات بازاریابی و فروش مورد استفاده قرار گیرد.

کلیدواژه‌ها: الگوریتم K-Means، داده‌کاوی، رفتار مشتری، فروشگاه آنلاین، خوشه‌بندی فازی.

۱- مقدمه

مدیریت ارتباط با مشتری به همه فرآیندها و فناوری‌هایی گفته می‌شود که در شرکت‌ها و سازمان‌ها برای شناسایی، طبقه‌بندی، ترغیب، گسترش، حفظ و ارایه خدمت به مشتریان به کار می‌رود [1]. سازمان‌ها با استفاده از *CRM* می‌توانند چرخه فروش را کوتاه‌تر و وفاداری مشتری و درآمد را افزایش دهند. سیستم مدیریت روابط با مشتری می‌تواند کمک کند تا مشتریان موجود حفظ شوند و مشتریان جدیدی جذب شوند. امروزه با گسترش سیستم‌های پایگاهی و حجم بالای داده‌های ذخیره شده در این سیستم‌ها، نیاز به ابزاری است تا بتوان داده‌های ذخیره شده را پردازش کرد و اطلاعات حاصل از این پردازش را در اختیار کاربران قرار داد. داده‌کاوی یکی از مهم‌ترین این روش‌ها است که به وسیله آن الگوهای مفید در داده‌ها با حداقل دخالت کاربران شناخته میشوند و اطلاعاتی را در اختیار کاربران و تحلیل‌گران قرار می‌دهند تا براساس آن‌ها تصمیمات مهم و حیاتی در سازمان‌ها اتخاذ شوند.

از آن‌جا که حجم داده‌های مشتریان بسیار زیاد است و بنابراین، برای این‌که بتوان مشتریان را طبقه‌بندی کرد از روش‌های داده‌کاوی برای تجزیه‌وتحلیل داده‌های مشتریان استفاده می‌شود و الگوریتم‌های داده‌کاوی مختلف برای تجزیه‌وتحلیل و طبقه‌بندی مشتریان توسط مجموعه داده‌های *UCI machin learning repository* بررسی می‌شوند [2].

پس از ورود داده‌ها به نرم افزار داده‌کاوی پیش پردازش به روی داده‌ها اعمال می‌شود برای این‌که تعداد ویژگی‌ها کاهش یابد تا به این وسیله میزان صحت و دقت الگوریتم‌ها بالا رود. سپس اثر آن را به‌روی الگوریتم‌های مورد نظر بررسی نماییم [3].

در بخش دوم ابتدا مدیریت ارتباط با مشتریان تعریف می‌شود، سپس داده‌کاوی و برخی از روش‌های آن مورد تجزیه‌وتحلیل قرار می‌گیرد و ارتباط بین داده‌کاوی و مدیریت ارتباط با مشتری بیان می‌گردد. در انتها نرم‌افزار رپید ماینر به‌عنوان محیط داده‌کاوی تعریف می‌گردد. روش‌های مختلفی برای بهبود ارتباط با مشتری از طریق خوشه‌بندی و طبقه‌بندی مشتریان مبتنی بر داده‌کاوی ارایه شده‌است [4]. در فصل دوم مروری برکارهای انجام شده در زمینه تجزیه‌وتحلیل و طبقه‌بندی مشتریان به وسیله روش‌های داده‌کاوی انجام شد [5].

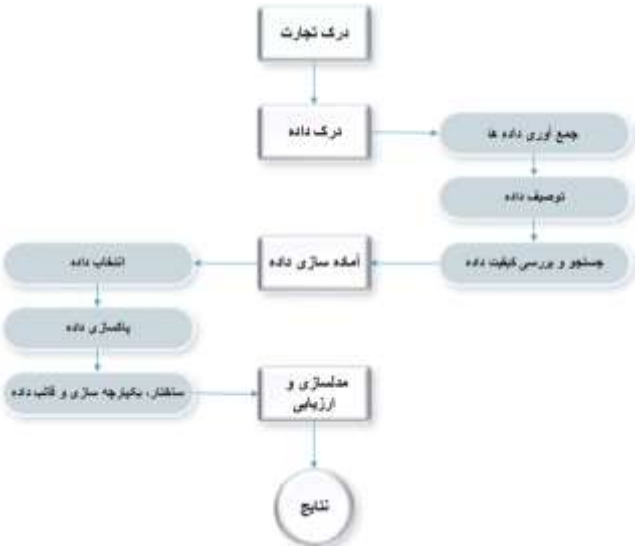
درسال ۲۰۱۷ میلادی، محققان به منظور شناسایی مشتریان بانک ملت و تدوین استراتژی‌های مناسب برای برخورد با آن‌ها از داده‌کاوی و ابزار آن مانند الگوریتم ژنتیکو الگوریتم میانگین k استفاده کردند.

۱. گروهی دیگر از محققان در سال ۲۰۱۹ از مدل طبقه‌کننده‌های ترکیبی استفاده کردند.
۲. گروهی از محققان در سال ۲۰۱۸ با استفاده از مدل شبکه عصبی برای اعتبار سنجی مشتریان استفاده کردند.
۳. بحث اعتبارسنجی مشتریان بانک توسط مدل‌ها یا اعتبارسنجی در تحقیقات اخیر بسیار به کار می‌رود. در سال ۲۰۰۸ محققان در مقاله‌هایی از مدل رتبه‌بندی تحلیل لینک با استفاده از ماشین بردار پشتیبان استفاده کردند.
۴. در سال ۲۰۱۸ با در نظر گرفتن سهم سود ایجاد شده، سود بالقوه و تعریف سودآوری مشتری یک مدل پیشنهاد کردند و مشتریان را بخش‌بندی نمودند [6].
۵. در برخی تحقیقات صورت گرفته در این حوزه، محققان در سال ۲۰۱۹ مشتریان یک بانک و یک کلوپ کتاب با استفاده از داده‌کاوی برای شناسایی مشتریان طبقه‌بندی شدند.

در بخش دوم تحقیق مجموعه داده مورد استفاده برای این تحقیق معرفی شد و خصوصیات و ویژگی‌های موجود در این مجموعه داده بیان گردید و روش پیشنهادی ارایه شد وبخش سوم مدل‌سازی بخش چهارم معیار ارزیابی و سپس در بخش پنجم مقایسه با سایر روش‌ها بیان شد و در بخش آخر نتیجه‌گیری گفته شد.

۲-روش تحقیق

مراحل داده‌کاوی بر اساس متدولوژی *Crisp* داده‌کاوی می‌باشد. ساختار کلی روش در **شکل ۱** آمده است.



شکل ۱- مراحل داده‌کاوی بر اساس متدولوژی Crisp [7].
Figure 1- Data mining steps based on the Crisp methodology.

۲-۱-درک تجارت

درک تجارت مرحله ابتدایی پژوهش می‌باشد که بر روی درک نیازمندی‌ها و اهداف پروژه از منظر تجارت تمرکز می‌کند. این دانش به یک تعریف مساله داده‌کاوی و یک طرح مقدماتی برای دستیابی به اهداف تبدیل می‌شود. در این پژوهش هدف این است که با به‌کارگیری از دانش داده‌کاوی با روش‌شناسی *Crisp* به کشف و استخراج مدل مفیدی بپردازیم و آن مدل را بهبود فروش انلاین مورد استفاده قرار دهیم.

۲-۲-درک و جمع‌آوری داده

مرحله‌ی دوم درک داده است که با جمع آوری داده اولیه شروع می‌شود و همچنین تشخیص مشکلات کیفیت داده، دریافتن درک داده و جست‌وجوی داده را در بر می‌گیرد.

۲-۳-آماده‌سازی داده

آماده‌سازی داده، سومین گام در متدولوژی کریسپ داده‌کاوی می‌باشد. این مرحله شامل انتخاب داده، پاک‌سازی داده، تعیین ساختار داده، یکپارچه سازی داده‌ها و قالب داده‌ها می‌باشد. این مرحله در داده‌کاوی بسیار مهم است زیرا پیش پردازش مناسب، در هر نوع پیش‌بینی در مدل‌سازی بسیار حیاتی می‌باشد [8].

۲-۴-انتخاب داده

در زمان جمع‌آوری داده‌ها از پایگاه داده، داده‌هایی را که برای این مطالعه ضروری بودند را انتخاب کردیم اما در این مرحله داده‌هایی را که دریافت کردیم، دوباره مورد بررسی قرار می‌دهیم.

۲-۵-پاک‌سازی داده

مرحله پاک‌سازی داده بخشی از آماده‌سازی داده می‌باشد و برای افزایش کیفیت داده به‌کار می‌رود. در مطالعه آنکیت آگراوال و همکاران [9] به منظور پاک‌سازی داده، رکوردها را کاهش داده و ویژگی‌هایی را که دارای مقادیر از دست رفته و یا دارای مقادیر تکراری می‌باشند، حذف کرده‌اند.

۲-۶-ساختار، یکپارچه‌سازی و قالب داده

یکی از مسایل مهم پیش پردازش این است که ساختار داده‌ها را طوری تنظیم کنیم که در هنگام مدل‌سازی دچار مشکل نشویم [10]. از آن‌جایی که اطلاعات جمع‌آوری شده در این مطالعه اغلب به صورت چند کلمه‌هایی هستند و حجم زیادی دارند، در هنگام کار با الگوریتم سرعت آن را کاهش می‌دهد، فرمت داده‌ها را از حالت کلمه‌ای به حروف اختصار تبدیل کردیم .

۲-۷-نرمال‌سازی داده ها

در مرحله پیش پردازش داده‌ها و جهت نرمال‌سازی از دو فیلتر *Normalize* استفاده می‌شود.

از فیلتر *Normalize* استفاده می‌کنیم که یک تکنیک پردازش است و برای تغییر اندازه‌ی مقادیر ویژگی در یک محدوده خاص مورد استفاده قرار می‌گیرد. نرمال‌سازی داده در هنگام برخورد با ویژگی‌های واحدهای مقیاسی مختلف بسیار مهم است.

۲-۸-مراحل خوشه بندی داده ها

فرایند کلی خوشه‌بندی داده‌ها شامل مراحل زیر است:

۱. نمایش الگو که معمولا شامل انتخاب و استخراج خصیصه است.
۲. تعریف معیار ارزیابی شباهت با توجه به دامنه داده‌ها.
۳. فرایند خوشه‌بندی یا گروه‌بندی.
٤. خلاصه‌سازی داده‌ها.
٥. در صورت نیاز اعتبارسنجی سیستم.

داده‌های مورد استفاده در این تحقیق از پایگاه داده‌هایی به نام *Online Shoppers Purchasing Intention* و از پایگاه داده‌ای *UCI machine learning repository* ، استخراج شده‌اند. مجموعه داده‌ها شامل ۱۰ ویژگی مطلق و ۸ ویژگی غیرمطلق است [11]. ویژگی درآمد می‌تواند به عنوان برچسب کلاس استفاده شود.

مدت‌زمان اجرایی، اطلاعات مربوط به محصول تعداد انواع مختلف صفحات بازدید شده توسط بازدیدکننده در آن جلسه و زمان کلی سپری شده در هر کدام از این صفحات را نشان می‌دهد. مقادیر این ویژگی‌ها از اطلاعات *URL* صفحات بازدید شده توسط کاربر گرفته می‌شوند و در زمان واقعی که یک کاربر یک عمل انجام می‌دهد، به طور مثال از یک صفحه به صفحه دیگر منتقل می‌شود. سرعت انتقال، سرعت خروج و ویژگی‌های ارزش صفحه ،معیارهای اندازه‌گیری شده توسط آنالیز گوگل برای هر صفحه در سایت تجارت الکترونیک را نشان می‌دهند [12]. سرعت انتقال برای صفحه وب به درصد بازدیدکنندگانی که از این صفحه وارد سایت می‌شوند و سپس انتقال بدون ایجاد هر درخواست دیگری برای سرور تجزیه‌وتحلیل در طول این جلسه اشاره می‌کند. سرعت خروج برای یک صفحه وب خاص برای مرور صفحه محاسبه می‌شود، درصدی که آخرین مورد در جلسه بود. ویژگی دیگری نشان‌دهنده مقدار میانگین برای یک صفحه وب است که یک کاربر قبل از تکمیل تراکنش های تجارت الکترونیک از آن بازدید می‌کند [13]. ویژگی روز ویژه نشان‌دهنده نزدیکی محل بازدید از زمان به یک روز خاص است. (به عنوان مثال روز مادر، روز ولنتاین) که در آن جلسات به احتمال بیشتری به تراکنش نهایی ختم می‌شوند. ارزش این ویژگی با توجه به پویایی تجارت الکترونیک مانند مدت‌زمان بین زمان سفارش و تاریخ تحویل تعیین می‌شود. این مجموعه داده‌ها شامل ویژگی‌ها و فیلدهای دیگری می‌باشد که مورد بررسی قرار می‌گیرد.

در این تحقیق بر روی مجموعه داده‌ها پس از پیش پردازش و آماده سازی، تحلیل آماری به منظور شناخت بیشتر داده‌ها انجام خواهد شد. پس از آن از الگوریتم‌های مختلف خوشه‌بندی با رویکرد نمونه‌گیری و ایجاد مجموعه داده‌های مختلف و رده‌بندی به منظور شناسایی ویژگی های تاثیرگذار بر فروش اینترنتی استفاده خواهد شد [14]. سپس انتخاب ویژگی‌های تاثیرگذار توسط روش الگویتم *FCM* در هر خوشه صورت می‌پذیرد.

۳-مدل‌سازی

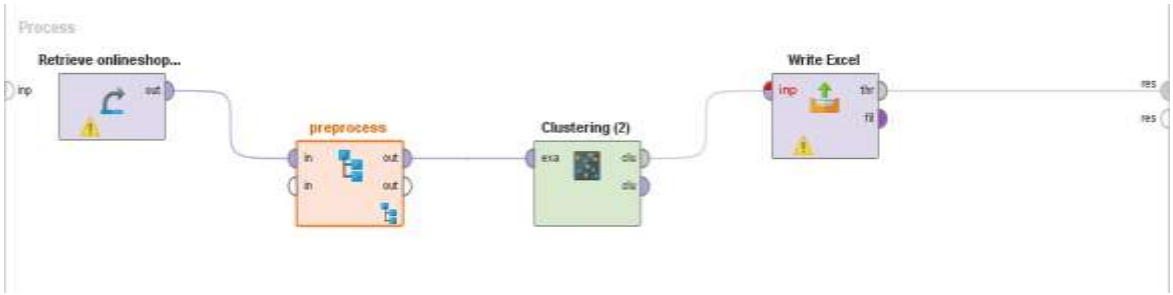
۱-۳-خوشه‌بندی

به منظور خوشه‌بندی رکوردها از الگوریتم k میانگین و برای پیدا کردن تعداد خوشه بهینه در این تحقیق از شاخص دیویس بولدین استفاده شده است. به این ترتیب که در هر مرحله تعداد خوشه های مختلف بر روی مجموعه داده تنظیم شده و سپس شاخص دیویس بولدین مدلسازی محاسبه شده است .همچنین تعداد خوشه‌ها در الگوریتم k میانگین ۲ خوشه تا ۹ تا خوشه تنظیم گردید .تعداد تکرار در خوشه‌بندی برابر ۱۰۰ و شاخص فاصله واگرایی برگمن بر اساس فاصله اقلیدسی تنظیم گردیده است. در جدول زیر مقدار شاخص دیویس بولدین برای هر تعداد از خوشه‌ها محاسبه شده‌است.

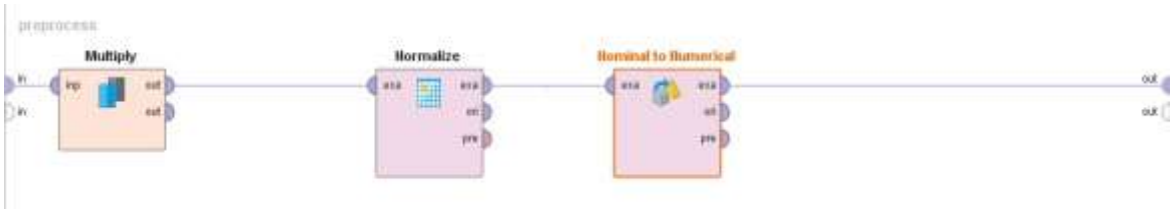
جدول۱- مقادیر شاخص دیویس بولدین برای تعداد خوشه‌های مختلف.
Table 1. Davis-Bouldin index values for different numbers of clusters.

تعداد خوشه	مقادیر شاخص دیویس بولدین
2	0.532
3	0.519
4	0.502
5	0.531
6	0.561
7	0.599
8	0.622
9	0.657

روش اعمال *kmeans* برمجموعه داده پس از اعمال فیلتر *Normalize* در **شکل ۲** مشاهده می‌شود.



شکل ۲- روش اعمال kmeans برمجموعه داده.
Figure 2- Method of applying kmeans to a data set.



شکل۳- پیش پردازش داده‌ها.
Figure 3- Data preprocessing.

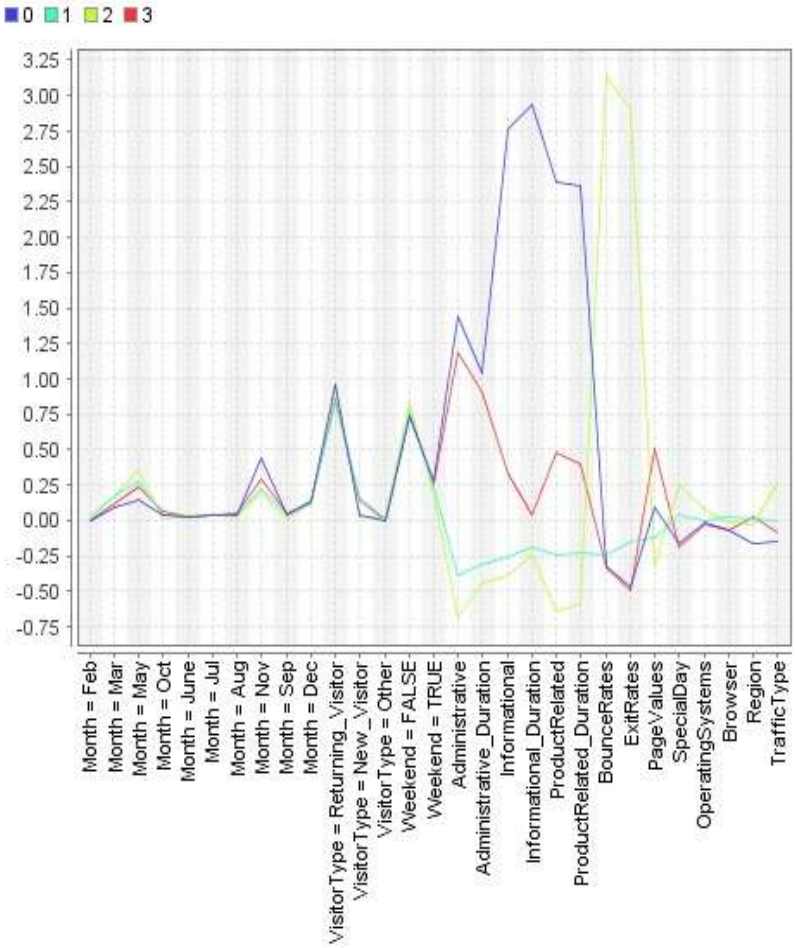
Cluster Model

```
Cluster 0: 587 items
Cluster 1: 8192 items
Cluster 2: 974 items
Cluster 3: 2577 items
Total number of items: 12330
```

شکل۴- نتیجه کلاستر کردن داده‌ها.
Figure 4- Result of clustering data.

جدول ۲- شاخص دیویس بولدین برای تعیین تعداد خوشه‌های بهینه در الگوریتم K-Means.
Table 2- Davis-Bouldin index for optimal cluster count in K-Means.

Attribute	Cluster_0	Cluster_۱	Cluster_۲	Cluster_۳
February	0.002	0.017	0.041	0.002
March	0.097	0.169	0.167	0.118
May	0.150	0.282	0.358	0.239
October	0.041	0.040	0.012	0.071
June	0.031	0.022	0.036	0.021
July	0.043	0.036	0.036	0.031
August	0.036	0.033	0.022	0.046
November	0.440	0.220	0.196	0.291
September	0.032	0.036	0.010	0.049
December	0.129	0.145	0.121	0.133
... VisitorType	0.966	0.837	0.954	0.853
... VisitorType	0.032	0.156	0.030	0.142
... VisitorType	0.002	0.007	0.016	0.004
= Weekend	0.727	0.767	0.835	0.752



شکل۵- نمودار حاصل از کلاستربندی داده‌ها.

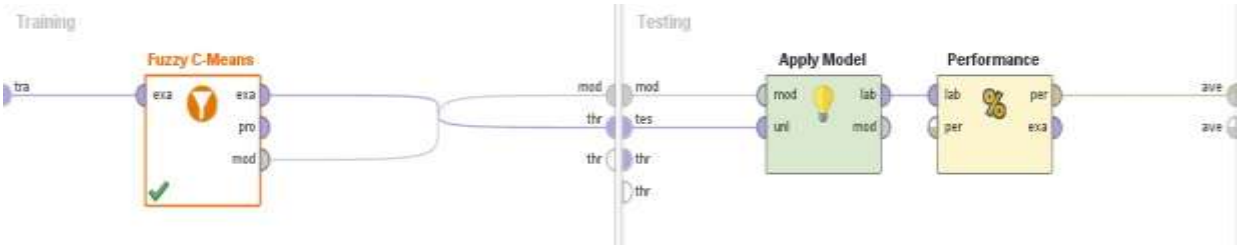
Figure 5- Graph resulting from data clustering.

سپس بعد از اعمال خوشه‌بندی ،داده‌های خوشه‌بندی شده را برای انتخاب ویژگی های تاثیر گذار به الگوریتم فازی *C_means* می‌دهیم.

۲-۳-انتخاب ویژگی‌های تاثیرگذار

به منظور افزایش دقت مدل‌سازی و استفاده بهینه از متغیرهای موجود در مجموعه داده از روش انتخاب ویژگی الگوریتم *FCM* به منظور انتخاب زیر مجموعه بهینه از مجموعه داده‌ها استفاده شده است.

روش اعمال *FCM* برمجموعه داده خوشه‌بندی شده توسط الگوریتم *kmeans* در **شکل ٦** مشاهده می‌شود.



شکل ۶- روش اعمال FCM برمجموعه داده.

Figure 6- Method of applying FCM to a data set.

ماتریس *CONFUSION* حاصل از اعمال *FCM* برمجموعه داده پس از اعمال الگوریتم *kmeans* در **جدول‌های ۲-۵** مشاهده می‌شود که درصد صحت طبقه‌بندی با *FCM* ۹۶٫۷۰% است.

۳-۳-معیارهای ارزیابی

یکی از مهم‌ترین قسمت‌هایی که در دسته بندی باید به آن توجه نمود، ارزیابی مدل دسته‌بندی است. در این بخش معیارهای ارزیابی برای دسته‌بندی مورد بررسی قرار خواهند گرفت. ارزیابی یک مدل دسته‌بندی میتواند بر اساس نمونه های آموزشی و آزمایشی صورت گیرد. برای ارزیابی باید برچسبی که مدل دسته‌بندی به آن دسته حمله نسبت داده شده، مقایسه شود. وقوع حالات مختلف برای دسته‌ها با توجه به مجموعه داده‌های ورودی برای دسته‌بندی با مقادیر *TP, FP, TN, FN* برای دو دسته مثبت و منفی در جدول زیر نشان داده شده است:

جدول ۳- وقوع حالات مختلف برای دسته‌ها.
Table 3. Occurrence of different states for groups.

		کلاس پیش‌بینی شده	
		Class=Yes	Class=No
کلاس واقعی	Class=Yes	TP	FN
	Class=No	FP	TN

در سیستم تشخیص معرفی شده ، داده‌ها به ۲ کلاس سرطانی و غیرسرطانی تقسیم‌بندی می‌شوند. ارتباط بین کلاس‌های واقعی وکلاس‌های پیش بینی شده در **جدول ۱** که به *confusion metrix* معروف می‌باشد، نشان داده شده است.

مهم‌ترین معیار برای تعیین کارایی یک الگوریتم، دسته‌بندی معیار *Accuracy* می‌باشد. این معیار دقت کل یک دسته‌بندی را محاسبه می‌کند. این معیارنشان دهنده این موضوع است که چند درصد از کل مجموعه داده‌ها به درستی دسته‌بندی شده است. رابطه زیر نحوه محاسبه معیار درستی را نشان می‌دهد.

$$ACCURACY = \frac{TP + TH}{TP + TN + FP + FN}. \quad (۱)$$

دو مقدار *TP* و *TN* مهم‌ترین مقادیری هستند که باید بیشینه شوند تا کارایی دسته‌بندی به حداکثر برسد. معیار خطای دسته‌بندی نیز از رابطه زیر به‌دست می‌آید. این رابطه دقیقا برعکس معیار *Accuracy* می‌باشد. کمترین مقدار برابر صفر (بهترین کارایی) و بیش‌ترین مقدار برابر یک (کم‌ترین کارایی) می‌باشد.

$$ER \equiv \frac{FN + FP}{TP + FP + FN + TN} = ۱ - Accuracy. \quad (۲)$$

معیارهای *Precision* درصدی را نشان می‌دهند که از میان تمامی دسته‌ها که توسر دسته بند به آن دسته نسبت داده شده اند، درست دسته‌بندی شده اند. به عبارتی دقت دسته‌بندی دسته i را با توجه به کل مواردی نشان می‌دهد که برچسبی i برای نمونه مورد بررسی توسر دسته بند پیشنهاد شده است. نحوه محاسبه این معیار در رابطه زیر نشان داده شده است. اندیس i در این پارامترها به این مفهوم است که پارامترها باید برای هر دسته i محاسبه شوند.

$$precision = \frac{TP}{TP + FP}. \quad (۳)$$

معیار *Recall* برای یک دسته، که از میان تمامی دسته های حملات متعلق به آن دسته، به درستی دسته‌بندی شده است. به عبارتی دقت دسته‌بندی دسته i را با توجه به کل نمونه های با برچسبی i نشان می‌دهد. نحوه محاسبه این معیار در رابطه زیر نشان داده شده است.

$$recall = \frac{\mathbb{TP}}{\mathbb{TP} + \mathbb{FN}}. \quad (۴)$$

نکته قابل توجه این است که معیار *Recall* کارایی دسته بند را با توجه به تعداد رخداد دسته i نشان می‌دهد درحالی که معیار *Precision* اساسا مبتنی بردقت پیشبینی دسته می‌باشد و بیان‌گر آن است که به چه میزان می‌توانیم به خروجی دسته بند اعتماد کنیم. معیار *F-measure* از ترکیبی معیارهای *Recall* و *Precision* به‌دست می‌آید و درمواردی استفاده می‌شود که نتوان اهمیت ویژه ای را برای هر یک از دو معیار *Recall* و *Precision* نسبت به یک‌دیگر قایل شد. رابطه زیر نحوه محاسبه این معیار را نشان می‌دهد.

$$F - MESURE - \frac{۲IP}{۲IP + FN + FP} \quad (۵)$$

جدول ۴- ماتریس CONFUSION حاصل ازطبقه‌بندی با FCM.

Table 4. CONFUSION matrix resulting from classification with FCM.

class precision	true 0.39084277952612284	true -0.3964618229174476	
97.92%	27	1271	pred. -0.3964618229174476
84.13%	106	20	pred. 0.39084277952612284
	79.70%	98.45%	class recall
class precision	true 0.39084277952612284	true -0.3964618229174476	

accuracy: 96.70%

ماتریس *precision* حاصل از اعمال *FCM* برمجموعه داده پس از اعمال الگوریتم *kmeans* در **جدول ۲** مشاهده می‌شود.مشاهده می‌شود که درصد *precision* خوشه‌بندی با *FCM* ۸۴٫۱۳% (*positive class*:۰٫۳۹۰۸۴۲۷۷۹۵۲۶۱۲۲۸۴) است.

ماتریس *recall* حاصل از اعمال *FCM* برمجموعه داده پس از اعمال الگوریتم *kmeans* در **جدول‌های ۵-۴** مشاهده می‌شود. مشاهده می‌شود که درصد *recall* طبقه‌بندی با *FCM* ۷۹٫۷۰% (*positive class*۰٫۳۹۰۸۴۲۷۷۹۵۲۶۱۲۲۸۴) است.

و همچنین *specificity* آن ۹۸٫۴۵% (*positive class*:۰٫۳۹۰۸۴۲۷۷۹۵۲۶۱۲۲۸۴) و ضریب کاپا ۰٫۸۰۰ و *classification_error* ۳٫۳۰% می‌باشد

۵-مقایسه

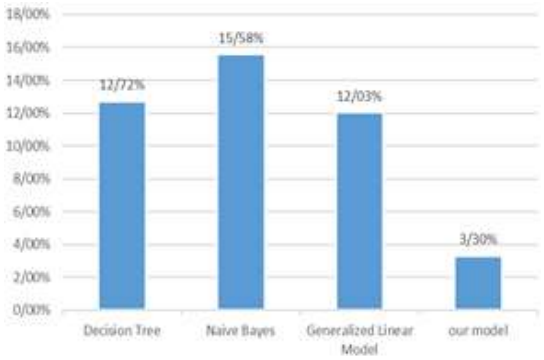
در این قسمت میزان صحت به‌دست آمده با چند الگوریتم دیگر مقایسه می‌شوند.

جدول Error! No text of specified style in document.-۵ مقایسه با سایر الگوریتم‌ها.

Table 5. Comparison with other algorithms.

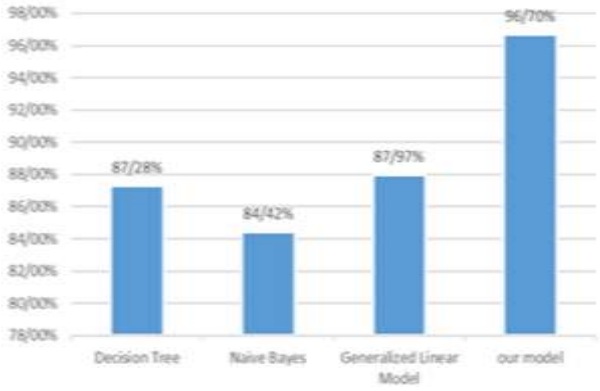
Model	Accuracy	Classification Error	Model Building Time
Decision tree	87.28%	12.72%	1.941 s
Naive bayes	84.42%	15.58%	1.182 s
Generalized linear model	87.97%	12.03%	1.316 s
Our model	96.70%	3.30%	1.0s

مقایسه این طبقه‌بندها در شکل‌های زیر نیز به صورت نمودار مشاهده می‌شود.



شکل۷-مقایسه درصد خطا.

Figure 7- Comparison of error percentage.



شکل۸- مقایسه درصد صحت.

Figure 8- Comparison of accuracy percentages.

۶- نتیجه

در این پژوهش به منظور تجزیه وتحلیل داده‌ها از دو نوع تجزیه‌وتحلیل استفاده خواهد شد:

۱. تجزیه‌وتحلیل توصیفی: در این نوع تجزیه و تحلیل، پژوهشگر داده‌های جمع‌آوری شده را با استفاده از شاخص‌های آمار توصیفی خلاصه و طبقه‌بندی می‌کند.
۲. تجزیه‌وتحلیل روابط: در این نوع تجزیه وتحلیل، روابط بین متغیرهای مستقل و وابسته مورد بحث و بررسی قرار می‌گیرد و پژوهشگر با رد یا تایید فرضیه‌های آماری به تایید یا رد رابطه بین متغیرها می‌پردازد.

مراحل داده‌کاوی بر اساس متدولوژی *Crisp* داده‌کاوی می‌باشد. دراین روش از ترکیب الگوریتم *kmeans* با خوشه‌بندی *FCM* استفاده شد. از مقایسه درصد صحت به‌دست آمده از خوشه‌بندی‌های موردنظر در این روش به این نتیجه می‌رسیم که میزان صحت الگوریتم پیشنهادی ما با ۹۶٫۷۰% نسبت به الگوریتم‌های دیگر بهتر است. به منظور توسعه نتایج حاصل از این تحقیقات می‌توان پیشنهادات ذیل را ارایه نمود:

۱. استفاده از روش های رده بندی جمعی به منظور افزایش دقت مدل‌سازی.
۲. پیاده‌سازی مدل پیشنهادی در این تحقیق بر روی مجموعه داده‌های دیگر و گزارش نتایج به‌دست آمده.
۳. استفاده از متغیرهای بیشتر به منظور مدل‌سازی بر روی مجموعه داده‌ها.
٤. روش‌های پیشنهاد شده در این تحقیق ،می‌تواند بر روی داده‌های دیگر استفاده شوند.

در تحقیقات آتی می‌توان با استفاده از پایگاه داده‌های بزرگ‌تر، تعداد رکوردها و دقت الگوریتم ها را افزایش داد [15].

منابع

[1] Makinde, A. S., Vincent, O. R., Akinwale, A. T., Oguntuase, A., & Acheme, I. D. (2020). An improved customer relationship management model for business-to-business e-commerce using genetic-based data mining process. *2020 international conference in mathematics, computer engineering and computer science (ICMCECS)* (pp. 1–7). IEEE. https://doi.org/10.1109/ICMCECS47690.2020.240875

[2] Lee, E. B., Kim, J., & Lee, S. G. (2017). Predicting customer churn in mobile industry using data mining technology. *Industrial management & data systems*, 117(1), 90–109. https://doi.org/10.1108/IMDS-12-2015-0509

[3] Jamalian, E., & Foukerdi, R. (2018). A hybrid data mining method for customer churn prediction. *Engineering, technology & applied science research*, 8(3), 2991–2997. http://dx.doi.org/10.48084/etasr.2108

[4] Ahmed, B. S., Maâti, M. L. Ben, & Al-Sarem, M. (2020). Predictive data mining model for electronic customer relationship management intelligence. *International journal of business intelligence research (IJBIR)*, 11(2), 1–10. https://doi.org/10.4018/IJBIR.2020070101%0A

[5] Van Nguyen, T., Zhou, L., Chong, A. Y. L., Li, B., & Pu, X. (2020). Predicting customer demand for remanufactured products: A data-mining approach. *European journal of operational research*, 281(3), 543–558. https://doi.org/10.1016/j.ejor.2019.08.015

[6] Kaur, E. R., & Kaur, E. K. (2017). Data mining on customer segmentation: A review. I*nternational journal of advanced research in computer science*, 8(5). https://doi.org/10.26483/ijarcs.v8i5.3479

[7] Joseph, S. I. T., & Thanakumar, I. (2019). Survey of data mining algorithm's for intelligent computing system. *Journal of trends in computer science and smart technology (TCSST),* 1(01), 14–24. https://doi.org/10.36548/jtcsst.2019.1.002

[8] Barreau, B., Carlier, L., & Challet, D. (2021). Deep prediction of investor interest: A supervised clustering approach. Algorithmic finance, 8(3–4), 77–89. https://doi.org/10.3233/AF-200296

[9] Agrawal, A., & Choudhary, A. (2016). Perspective: Materials informatics and big data: Realization of the "fourth paradigm" of science in materials science. *APL materials,* 4, 53208. https://doi.org/10.1063/1.4946894

[10] Halkin, A. (2018). Emotional state of consumer in the urban purchase: processing data. *Foundations of Management*, *10*(1), 99-112. https://doi.org/10.2478/fman-2018-0009

[11] Dananjali, K. T., Ekanayake, J. B., & Karunaratne, A. S. (2019). K-means clustering algorithm to predict the badulla tomato price based on weather factors. *International Research Conference of Uva Wellassa University*. Uwa Wellassa University of Sri Lanka. https://www.researchgate.net/publication/331375615

[12] Sari, P. K., & Purwadinata, A. (2019). Analysis characteristics of car sales in E-commerce data using clustering model. *Journal of data science and its applications*, 2(1), 19–28. https://doi.org/10.21108/JDSA.2019.2.19%0D

[13] Madimutsa, T. (2017). *Data mining for automated user's behaviour classification for an e-commerce system [Thesis].*

[14] Hasan, F., & Mia, M. J. (2019). Association rules and clustering on sparse data of a leading online retailer. *International journal of computer science and information security (IJCSIS)*, 17(4). https://www.researchgate.net/publication/333132169

[15] Larasati, D. A. H. D., & Sutrisno, T. (2018). *Tourism site recommendation in jakarta using decision tree method based on web review*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3268964