تصميم خوارزمية كشف الاحتيال في مواقع التجارة الإلكترونية بهدف تحسين الدقة والسرعة في عمليات الكشف

الطالب: علي إبراهيم

إشراف:د.بسيم عمران

ملخص

تعتبر أنظمة الكشف عن الاحتيال من أهم الأنظمة المستخدمة في مواقع التجارة الإلكترونية، وذلك بهدف حماية الموقع والمستهلك والشركات التي تعرض منتجاتها عبر مواقع التجارة الإلكترونية من الهجمات الاحتيالية، وعليه تقوم معظم مواقع التجارة الإلكترونية العالمية بتخصيص مبالغ مادية كبيرة بهدف تحسين وتطوير بنية أنظمة كشف الاحتيال لديها بهدف جعل النظام أكثر دقة و كفاءة.ولكن على الرغم من التطور الكبير في مجال الكشف عن الاحتيال إلا أن أغلب الحلول المقدمة ما تزال تعاني من مشكلتي الدقة والسرعة في التنفيذ، وعليه قام الباحث بتقديم نظام لكشف الاحتيال يعتمد بشكل أساسي على دراسة سلوك المستخدم ضمن الموقع، وتطبيق الحل على مجموعة بيانات Kaggle.com على مرحلتين ، الأولى تطبيق الحل على قاعدة البيانات الأساسية ، والثانية تطبيق الحل على قاعد البيانات الموسعة حيث قام الباحث بزيادة عدد السجلات وقام بتقديم نتائج الدقة ومقارنتها مع نتائج خوارزميات التنقيب عن المعرفة في قواعد البيانات المعرفية والتي تستخدم بشكل كبير في مجال الكشف عن الاحتيال في مواقع التجارة الإلكترونية.تقدم هذه المقالة حلاً ديناميكياً للاحتيال يعتمد على دراسة سلوك المستخدم ضمن موقع التجارة الإلكترونية، ومن ثم مقارنة

نتائج الحل المقترح من قبل الباحث مع الحلول الثابتة المعتمدة على خوارزميات تقنيات التنقيب عن المعرفة في قواعد البيانات المعرفية، كما توضح المقالة أثر أخذ الخصائص المناسبة وإضافتها على مجموعة البيانات Kaggle (وهي جدول يحتوي على مناقلات خاصة بموقع تجارة إلكترونية)، حيث تم دراسة أثر أخذ الخصائص المناسبة بعين الاعتبار وأثرها على الدقة مع ازدياد عدد السجلات(عدد المناقلات الخاصة بكل مستخدم ضمن الموقع)، كما قام الباحث بتقديم هيكلية تقوم بجمع المعلومات المطلوبة عن تفاعل المستخدم مع الموقع، ومن ثم مقارنة السلوك الحالي للمستخدم مع السلوك السابق المخزن ضمن قاعدة البيانات لتحديد هل المناقلة الحالية شرعية أم أنها عملية احتيالية.

Abstract:

Fraud detection system is one of the most important systems used in e-commerce sites، which protect e-commerce website، consumers and companies from fraudulent attacks .Most e-commerce website spent a lot of money in improving the architecture of its fraud detection system to make it more accurate and efficient. However، despite the great development in the field of fraud detection، most of the solutions presented still have accuracy and speed problems، therefore the researcher presented a fraud detection system that relies primarily on studying user behavior within the site in two stages ,First stage is applying the solution to the primary database، second stage is applying the solution to the database after increasing the number of records، after that researcher presents accuracy and efficiency of system. This article presents dynamic solution for fraud depending on studying users' behavior in E-commerce website and comparing results of suggestion solution with other static results of KDD's

algorithms. This article explains the benefits of adding required attributes to kaggle dataset and studies the benefits of adding required attributes at accuracy of fraud detection in e-commerce website.Within the articles, The researcher provides structure of fraud detection system which collect data about user's interaction with e-commerce website and compare the current user's behavior with previous stored user's behavior to determine if the current transaction is fraud or not.

مقدمة

ازداد في الآونة الأخيرة استخدام أنظمة الكشف عن الاحتيال بشكل واسع وكبير، ليس فقط في مواقع التجارة الإلكترونية وانما تعداها الى استخدامات واسعة في مواقع التواصل الاجتماعي، أنظمة البورصة العالمية وغيرها ،وذلك بسبب الأثر السلبي والخسائر المادية الكبيرة التي تسببها مثل هذه النوعية من عمليه النصب الإلكتروني، إن الحد من الخسائر التي تتكبدها مواقع التجارة الإلكترونية سنوياً تعتبر أهم الأهداف لأنظمة كشف الاحتيال، إلا أن مثل هذه الأنظمة وعلى الرغم من التطورات الكبيرة التي طرأت عليها لا تزال تعاني من العديد من المشاكل نذكر[1] منها:

1- **مشكله الدقة:** تعتبر دقه أنظمة الكشف الاحتيالي لتحديد هل المناقلة الحالية شرعيه أم احتياليه العامل الأساسي لتقييم مثل هذه الأنظمة، وخاصة أن دقة مثل هذه الأنظمة تتأثر بشكل كبير بعدد السجلات، حيث ما تزال تعاني الأنظمة المعتمدة على تقنيات التنقيب عن المعرفة في قواعد البيانات المعرفية من مشكلة انخفاض الدقة بزيادة عدد السجلات وزيادة حجم قاعدة البيانات الخاصة بمواقع التجارة الإلكترونية وذلك في حال عدم أخذ الخصائص المناسبة بعين الاعتبار، وخاصة تلك الخصائص التي ترتبط ارتباطاً وثيقاً بسلوك المستخدم ضمن موقع التجارة الإلكترونية وتفاعلاته ضمن الموقع، حيث تقدم خوارزميات التنقيب عن المعرفة في قواعد البيانات المعرفية حلولاً ثابتة ستاتيكية لمشكلة الاحتيال والتي تتأثر دقتها بشكل كبير مع ازدياد عدد السجلات ضمن قاعدة البيانات .[2]

2- **مشكلة الأداء:** تعتمد بنية أغلب أنظمة الكشف عن الاحتيال في مواقع التجارة الإلكترونية المعتمدة على تقنيات التنقيب عن المعرفة في قواعد البيانات المعرفية على عملية مسح كامل لقاعدة البيانات ، وذلك لاتخاذ القرار المناسب هل المناقلة الحالية شرعية أم أنها عملية احتيالية الأمر الذي سيكلف الكثير من الوقت واستهلاك الموارد المادية للنظام، وخاصة أن قاعدة البيانات الخاصة بمواقع التجارة الإلكترونية تتزايد بشكل أسي بمرور الوقت.[3]

قام الباحث من خلال هذه المقالة بتصميم نظام لكشف الاحتيال في مواقع التجارة الإلكترونية بالاعتماد على دراسة سلوك المستخدم ضمن الموقع، وذلك من خلال قيام الباحث بدراسة وإضافة العديد من الخصائص التي ترتبط ارتباطاً وثيقاً بسلوك المستخدم، حيث يوجد العديد من هذه الخصائص لم تكن موجودة في الدراسات الخاصة بأنظمة الكشف عن الاحتيال ، ولكن الباحث اعتمد بشكل أساسي على أنظمة تحليل البيانات، وتحليل المحتوى، واستنتاج أهم الخصائص التي ترتبط بسلوك المستخدم ضمن الموقع ، كما وقام الباحث بعملية تحسين للأداء من خلال توزيع عملية جمع المعلومات المطلوبة على كامل النظام، وذلك عكس باقي الحلول المطروحة المعتمدة على تقنيات التنقيب عن المعرفة في قواعد البيانات المعرفية والتي تعتمد بشكل أساسي على جمع كل عمليات المقارنة وعمليات جمع المعلومات ضمن مرحلة واحدة ، الأمر الذي سيؤثر بشكل أساسي على زمن الانتظار الخاص بالمستخدم وذلك لتحديد هل المناقلة الحالية شرعية أم أنها عملية احتيالية. اعتمد الباحث بعد عملية جمع المعلومات عن المستخدم القيام بمقارنة سلوك المستخدم الحالي مع السلوك السابق له والمخزن ضمن قاعدة البيانات لتحديد هل المناقلة الحالية شرعية أم أنها عملية احتيالية، ومن ثم قام بدراسة دقة النظام المقدم ومقارنته مع أكثر خوارزميات كشف الاحتيال استخداماً وذلك باستخدام معايير الدقة الخاصة بأنظمة كشف الاحتيال، بحيث يحقق هذا النظام المقترح أعلى دقة وبأفضل أداء.

مشكلة البحث

تعاني أغلب الحلول المقدمة لحل مشكلة الاحتيال في مواقع التجارة الإلكترونية من مشكلة الدقة بمرور الزمن والمشكلة الثانية مشكلة الأداء أو سرعة التنفيذ(زمن الكشف عن الاحتيال) نتيجة اعتماد

أغلب الحلول المقدمة من قبل تقنيات التنقيب عن المعرفة في قواعد البيانات المعرفية على مسح كامل لقاعدة البيانات وذلك بهدف الحصول على نتائج دقيقة، علماً بأن حجم قواعد البيانات الخاصة بمواقع التجارة الإلكترونية تتزايد بشكل أسي بمرور الزمن الأمر الذي سيسبب مشكلة للحلول التي تعتمد على مسح كامل لقاعدة البيانات.

أهداف البحث

يهدف البحث إلى تقديم نظام لكشف الاحتيال يعتمد على تتبع سلوك المستخدم مع الموقع من خلال تسجيل تفاعلات المستخدم مع الموقع وتخزين أهم الخصائص الخاصة بالمستخدم وتفاعلاته ضمن موقع التجارة الإلكترونية بهدف الوصول إلى نظام يتميز بالدقة الجيدة وزمن الكشف الجيد (والمقصود بالدقة هو عدد المناقلات الاحتيالية التي تم كشفها من قبل النظام والمقصود بزمن الكشف هو الزمن اللازم للنظام لتحديد هل المناقلة الحالية شرعية أم أنها عملية احتيالية) ، كما يهدف البحث إلى المقارنة بين الحلول الديناميكية التي تعتمد على سلوك المستخدم ومقارنتها مع الحلول الثابتة المقدمة من تقنيات التنقيب عن المعرفة في قواعد البيانات المعرفية، حيث اعتمد الباحث على أخذ أهم الخصائص الخاصة بسلوك المستخدم ضمن موقع التجارة الإلكترونية اعتماداً على دراسات سابقة ، وإضافة خصائص جديدة مهمة ولها أثر في تحديد سلوك المستخدم بطريقة أكثر دقة وشمولية سيتم تحديدها في المقالة.

أهمية البحث

تأتي أهمية البحث من كون الحصـول على نظام دقيق وذو كفاءة عالية تتزايد بمرور الوقت نتيجة اعتماد النظام المقترح على سلوك المستخدم يعتبر حل جيد وذو دقة متزايدة بمرور الوقت.

منهج وفرضيات البحث

يعتمد هذا البحث على المنهج التطبيقي وذلك من خلال تصـميم نظام لكشـف الاحتيال وتطبيقه على مجموعة البيانات العالمية Kaggle[4] الأسـاسـية والموسـعة، والحصـول على نتائج الدقة باسـتخدام برنامجي Weka,RStutio وتقديم نتائج الدقة من خلال معايير الدقة المسـتخدمة في تقييم أنظمة كشف الاحتيال.

مجموعة البيانات Kaggle الأساسية والموسعة:

يقدم موقع kaggle.com عدد كبير من مجموعات البيانات الخاصة بمواقع التجارة الإلكترونية والتي تحتوي على جداول مناقلات خاصة بالمستخدمين ضمن الموقع، حيث قام الباحث باختيار مجموعة بيانات تحتوي هذه المجموعة على 284808 سجل خاصة بالمناقلات بين المستخدمين وموقع التجارة الإلكترونية، وتحتوي على 22 خاصية من الخاصيات من بينها المنتجات التي قام المستخدم بشرائها، يضاف لها الخصائص التالية:

1- الزمن Time: وهو الوقت الذي استغرقه المستخدم للقيام بعمليات الشراء عبر موقع التجارة الإلكترونية.
2- الكمية Amount: وهي كمية المنتجات التي قام بها صاحب البطاقة.
3- الصنف Class: هل هذه المناقلة هي مناقلة شرعية أم أنها عملية احتيالية.

حيث تم أخذ عينات من قاعدة البيانات التي تم جمعها ابتداءً من عام 2013 حتى عام 2022، وهي تحوي على عمود أخير يحدد فيما إذا كانت المناقلة الحالية شرعية أم لا، بغية مقارنة نتائج الحلول المقدمة والخوارزميات التي تتم تطبيقها على قاعدة البيانات واختبار الدقة. قام الباحث لجعل عملية التحليل واختبار الدقة الخاصة بعمليات المقارنة بين الحل المقترح من قبله والحلول السابقة بتوسعة قاعدة البيانات وذلك من خلال زيادة عدد السجلات المدروسة إلى 532251 من خلال دمج مجموعة بيانات أخرى لموقع Kaggle بنفس الخصائص ، ومن ثم قام الباحث بزيادة عدد الخصائص المدروسة والخاصة بسلوك المستخدم ضمن الموقع وذلك ضمن جدولين هما Complete_User_Behavior و User_profile والتي بعضها لم يكن موجود في دراسات سابقة (سيتم شرح هذين الجدولين في معرض المقالة)، واختبار الحل المقترح على قاعدة بيانات Kaggle الموسعة والمعدلة من قبل الباحث.

حيث قام الباحث بالاعتماد على محموعة البيانات kaggle وذلك لكونها مجموعة بيانات متوازنة واتساقية وخالية من الضجيج حيث لا يوجد حاجة إلى إعادة معالجة البيانات، حيث تمت دراسة سلوك نحو 4112 مستخدم ، والجدير بالذكر أن مجموعة البيانات Kaggle تحتوي على عمود

أخير يحدد هل المناقلة الحالية شرعية أم أنها احتيالية، بقيمة =1 للحالات الشرعية وقيمة =0 للحالات الاحتيالية، الأمر الذي مكن الباحث من التأكد من دقة النتائج،حيث تقدر عدد الحالات الاحتيالية ضمن مجموعة البيانات الأساسية ب 6222 حالة احتيالية و 278586مناقلة شرعية، أما في قاعدة البيانات الموسعة يوجد لدينا نفس عدد المستخدمين ، بينما تقدر عدد الحالات الاحتيالية ب13225 مناقلة احتيالية و 519026 مناقلة شرعية.يوضح الجدول التالي بينة مجموعة البيانات المقدمة من قبل kaggle:

Trans ID	User ID	Time	Amount	Class	Product1	Amount1	….	transType

الجدول (1) يوضح بنية مجموعة البيانات الخاصة بالمناقلات المقدمة من موقع kaggle

كما يوضح الجدول التالي بينة الجدول الخاص بمعلومات المستخدمين ضمن الموقع :

User ID	User Firstnamre	Last Name	Age	Education	Credit Card ID	User Name	Password

الجدول (2) يوضح معلومات المستخدمين ضمن موقع kaggle

كما يوضح الجدول التالي المعلومات الخاصة ببطاقات الائتمان :

Credit ID	User _Id	Credit _Card_Num	Credit Card password	Credit Card_ Bank

الجدول (3) يوضح معلومات بطاقات الإئتمان ضمن موقع kaggle

كما سيتم توضيح جدولي خصائص المستخدمين المضافة من قبل الباحث في معرض المقالة.

معايير الدقة والاختبار:

توجد العديد من معايير الدقة التي تستخدم في عمليات المقارنة بين خوارزميات الكشف عن الاحتيال في مواقع التجارة الإلكترونية، ولكن تم اعتماد أكثر هذه المعايير جودة وكفاءة وهي [5]:

1- TP Rate.

2- FP Rate.

3- $precision\ P = TP/(TP + FP)$

4- $Recall\ R = TP/(TP + FN)$

5- $F1 = 2 * Precision * Recall/(Precision + Recall)$

6- MCC.

7- Roc Area.

8- PRC Area.

أهم الدراسات المرجعية السابقة:

تم تقديم العديد من الأنظمة الخاصة بعمليات الكشف عن الاحتيال في مواقع التجارة الإلكترونية، نذكر منها:

1- RAPTIDAR,R, 2021 [6] : قام Fraud Detection using GA and AI الباحث بالدمج بين الشبكات العصبونية والخوارزميات الجينية وقسم الحل على ثلاث مراحل حيث تم في المرحلة الأولى تدريب الشبكة العصبونية كما قام الباحث بالاعتماد في المرحلة الثانية على خوارزمية SVM لزيادة الدقة، وطبق خرج خوارزمية SVM على الخوارزمية الجينية كمرحلة ثالثة، واختبر دقة الحل على قاعدة بيانات Kaggl، وعلى الرغم من الدقة التي تم الحصول عليها(95.923%) إلا أن النظام يتميز بالتعقيد وزمن التنفيذ العال، كونه اعتمد على ثلاثة مراحل للوصول للحل إضافة إلى الصعوبة في التطوير والتعديل على الحل، حيث اعتمد الباحث على دراسة عدد الحالات الاحتيالية

المكتشفة كمعيار أساسي للدقة(لم يتطرق الباحث ضمن المقالة إلى زمن التنفيذ أو زمن الكشف عن الاحتيال).

2- Keveort,A -2018. Faud Dtection Using Decision tree and Smote Decision Tree[7]قام الباحث بتقديم نظام يعتمد على خوارزمية شجرة دعم القرار وشجرة دعم القرار المعدلة وقام بتطبيق الحل على قاعدة البياناتkaggle، حيث قسم الباحث الحل إلى مرحلتين : الأولى تعتمد على تطبيق خوارزمية شجرة أخذ القرار المعدلة، ومن ثم تطبيق خوارزمية شجرة دعم القرار وكانت دقة النتائج أقل من الحل السابق(94.281%) (حيث اعتمد في قياس الدقة على دراسة عدد حالات الاحتيال المكتشفة)، وبزيادة عدد السجلات سنحصل على دقة أقل وسيكون سرعة التنفيذ أكبر وبالتالي الحل غير مفيد في حال قواعد البيانات الضخمة بسبب اعتماد الباحث على خوارزميات ثابتة ولم يأخذ كامل الخصائص المناسبة ضمن الدراسة الخاصة بسلوك أو اهتمامات المستخدم بعين الاعتبار والتي تعتبر المقياس الأهم في تقييم هل المناقلة الحالية شرعية أم أنها عملية احتيالية.

3- Holland,J,2021- Using logistic Regression and KNN to detect fraud Transaction in e-Commerce: حيث قام الباحث بتقديم حل يعتمد على الدمج بين خوارزمية logistic Regression وخوارزمية الجار الأقرب [8] ، حيث اعتمد الباحث في المرحلة الأولى على استخدام خوارزمية الجار الأقرب ثم طبق خرج هذه المرحلة كدخل لخوارزمية Logistic Regression ،حيث أعطت نتائج أفضل من تطبيق كل خوارزمية على حدا وكانت الدقة (96.341%)، ولكن الحل المقترح لم يأخذ كامل الخصائص الواجب دراستها بعين الاعتبار، كما أن الحل المقدم يتميز بالتعقيد وزمن التنفيذ العاليين وصعوبة التطوير والتعديل على هذا النموذج.

نلاحظ أن أغلب الحلول المقدمة سابقاً والأكثر انتشاراً واستخداماً تعتمد بشكل أساسي على الخوارزميات التالية[9]:

1- KNN for K=3 and K=7.

2- SVM.

3- Decision Tree.

4- Smote Decision Tree.

5- Logistic Regression.

6- Navie Bayes.

7- Smote Navie Bayes.

8- Random Forest.

وعليه قام الباحث بتطبيق هذه الخوارزميات على قاعدتي بيانات Kaggle الأساسية والموسعة، حيث اعتمد في عمليات المقارنة على برنامجي Weka, RStudio[10] وكانت النتائج على النحو التالي:

Algorithms	Accuracy		AVG
	Weka	Rstudio	
K=3،KNN	95.433%	95.323 %	95.378%
K=7،KNN	95.593%	95.455%	95.524 %
Navie Bayes	95.813%	95.772%	95.7925%
Logistic Regression	96.632%	96.451%	96.5415%
SVM	97.491%	97.421%	97.456%
Navie SMOTE	92.029%	92.3738%	92.18645%
Decision Tree	93.199%	93.0856%	93.14223%
Random Forest	92.71103%	92.997%	92.854015%
Decision Tree SMOTE	91.09377%	90.2627%	90.678235%

الجدول(4): يوضح نتائج مقارنة الخوارزميات على قاعدة البيانات الأساسية

نلاحظ من الدراسة السابقة أن أفضل الخوارزميات المقدمة في حل مشكلة كشف الاحتيال هي خوارزمية SVM بمعدل دقة متوسطة يقدر ب97.456% وأن أسوء الخوارزميات المقدمة هي خوارزمية Decision Tree Smote بمعدل دقة يصل إلى 90.678%.

المرحلة الثانية: قام الباحث بالمقارنة بين الخوارزميات السابقة بتطبيقها على قاعدة بيانات kaggle الموسعة بحيث أصبح عدد السجلات فيها 532251 وكانت النتائج على النحو التالي:

	Accuracy		AVG
Algorithms	Weka	Rstudio	
K=3،KNN	92.325%	91.9952%	92.1385%
K=7،KNN	92.5598%	92.137%	92.3484%
Navie Bayes	91.926%	91.0725%	91.499%
Logistic Regression	94.6398%	94.1495%	94.3945%
SVM	95.599%	95.279%	95.4141%
Navie Bayes SMOTE	88.3139%	88.1379%	88.2259%
Decision Tree	84.951%	84.5426%	84.7468%
Random Forest	86.348%	85.42%	81.884%
Decision Tree SMOTE	81.74905%	80.969%	81.359025%

الجدول(5): يوضح نتائج مقارنة الخوارزميات على قاعدة البيانات المعدلة والموسعة من قبل الباحث

نلاحظ أنه بزيادة عدد السجلات فإن دقة الخوارزميات قد انخفضت بشكل كبير، وذلك بسبب عدم أخذ كل الخصائص المطلوبة بعين الاعتبار، الأمر الذي يؤثر بشكل كبير على درجة الدقة المطلوبة، حيث نلاحظ أن أكثر الخوارزميات تأثراً هي خوارزمية Decision Tree Smote وأقل الخوارزميات تأثراً هيٍ بمقدار 2.6% وعليه فإن دقة هذه الخوارزميات والتي تستخدم بشكل كبير في مجال كشف الاحتيال تتناقص بشكل ملحوظ كلما كبر حجم قاعدة البيانات لدينا، وبالتالي لحل هذه المشكلة،

ولجعل دقة النظام تزداد بمرور الوقت ، اقترح الباحث القيام بإضافة خصائص جديدة لقاعدة البيانات، هذه الخصائص مرتبطة بسلوك المستخدم ضمن قاعدة البيانات.

بنية النظام المقترح لحل مشكلة الاحتيال:

اعتمد الباحث في الحل المقترح على تسجيل سلوك المستخدم ضمن الموقع وتسجيل تفاعلات المستخدم ضمن جدولين:

الجدول الأول: هو Complete User behavior وهو يحتوي على المعلومات التالية:

Attribute	Description
User_Behavior_ID	رقم المستخدم ضمن الجدول
User_ID	رقم المستخدم من جدول المستخدمين
User_Type	نوع المستخدم وهو إما جديد أو يحتوي على عدد قليل من المناقلات أو يحتوي على عدد جيد من المناقلات
User_Name	اسم المستخدم
Average_Session_Time	المدة المتوسطة للجلسة
NPS(Net Promoter Score)	عدد النقرات التي يقوم بها المستخدم ضمن الجلسة
Summary_Reading	قراءة ملخص المنتج قبل شراءه
Reading_All_Summary	قراءة كل الملخصات قبل الشراء
Read_similar_product	قراءة منتجات مشابهة للمنتج الحالي
Speed_in_typing	سرعة المستخدم في الكتابة(وهي المتوسط الحسابي لعشر إدخالات للمستخدم في الجلسة الواحدة

amount_in_transaction	كمية الشراء التي يقوم بها المستخدم في الجلسة الواحدة هل هي كبيرة أو متوسطة أو قليلة
Changing_Ip	هل المستخدم يقوم بتغيير IP بشكل مستمر أم لا
Last_IP	آخر عنوان IP للمستخدم
Changing_Tanspotation_location	هل المستخدم يغير عنوان الشحن بشكل مستمر[11]
Direct_Purchase	هل المستخدم يقوم بالشراء مباشرة أم لا
Purchase_from_many_Credit_Cards	هل المستخدم يقوم بالشراء من أكثر من بطاقة ائتمانية
Required Discountig	هل المستخدم يطلب حسم أم لا
Times Between Transaction	الفترة الزمنية بين طلبي شراء(هل هنالك فترة زمنية بين طلبي شراء أم لا)
Purchase_Porduct	هل المستخدم يقوم بشراء منتجات ذات أسعار مرتفعة
Product_Type	هل المستخدم يقوم بشراء منتجات متنوعة في نفس الجلسة أم لا
Trasaction Type	هل المناقلة الحالية شرعية أم غير شرعية

الجدول(7) يوضح الخصائص المدروسة والمضافة من قبل الباحث

يضاف إلى الجدول السابق جدول ثان هو ملخص عن تفاعلات المستخدم ضمن الموقع وهو User Profile يتم فيها تسجيل وتعديل هذا السجل بشكل دوري بناء على تفاعلات المستخدم ، حيث يتم التعديل في حال كانت المناقلة الحالية للمستخدم شرعية، بحيث يتم أخذ متوسط حسابي لبعض الخصائص مثل: متوسط سرعة الكتابة، الفترة الزمنية بين جلستين، المدة الزمنية للجلسة، يضاف

إلى ما سبق التعديل على بعض الخصائص مثل: هل المستخدم قام بتغيير عنوان الشحن، هل المستخدم قام بتغيير IP البلد الذي يشتري أو يتصفح منها عادة، وغيرها من الأمور الأخرى.

حيث يتم إعطاء أوزان لهذه الخصائص ، بحيث يتم مقارنة سلوك المستخدم الحالي مع سلوك المستخدم ضمن قاعدة البيانات (السجل الموجود ضمن جدول User_Profile), ويتم اتخاذ القرار بناءً على قيم السلوك الحالي للمستخدم وهو:

1- فإذا كان وزن سجل المستخدم الحالي => حد العتبة وهو 7 من أصل عشرة تعتبر المناقلة الحالية مناقلة شرعية.
2- فإذا كان وزن سجل المستخدم الحالي < حد العتبة وهو 7 من أصل عشرة تعتبر المناقلة الحالية مناقلة احتيالية.

من الجدير بالذكر هنا أن النظام لدينا يبدأ بإعطاء نتائج ذات دقة مقبولة بعد قيام المستخدم ب 10 مناقلات مع الموقع سواء من ناحية الشراء أو التصفح أو غيرها من التفاعلات الأخرى.

كما يقدم البحث إضافة للحلول المطروحة التي تقوم بدراسة سلوك المستخدم وذلك من خلال ما يلي:

1- تم تسجيل كامل مناقلات المستخدم ضمن الموقع سواء أكانت المناقلة الحالية هي عملية شراء أو عملية تصفح أو أي عملية تفاعل بين المستخدم الحالي والموقع، الأمر الذي يسرع عملية الحصول على نتائج دقيقة من قبل النظام المطروح، حيث تقدر عدد المناقلات التي يقوم بها المستخدم وذلك للحصول على نتائج ذات دقة مقبولة بحوالي 10 مناقلات، وهذا الأمر يعتمد على مدى تفاعل المستخدم مع الموقع ضمن المناقلة الواحدة، هل قام بالكثير من التفاعلات أو كانت تفاعلاته قليلة.
2- قام الباحث بتوزيع عمليات جمع المعلومات على مستوى النظام ككل، وتسجيل كل تفاعل للمستخدم مع الموقع مع كل حركة يقوم بها المستخدم ضمن الموقع ووضع هذه المعلومات ضمن جدول Complete_User_Behavior، ومن ثم أخذ المتوسطات الحسابية لهذه التفاعلات وتسجيلها ضمن جدول User_profile وذلك لمقارنة السلوك الحالي للمستخدم

مع السلوك السابق والذي هو سجل واحد موجود في جدول User_profile الأمر الي يسرع عملية المقارنة ويحسن من زمن الكشف عن المناقلة الحالية هل هي شرعية أم أنها عملية احتيالية.

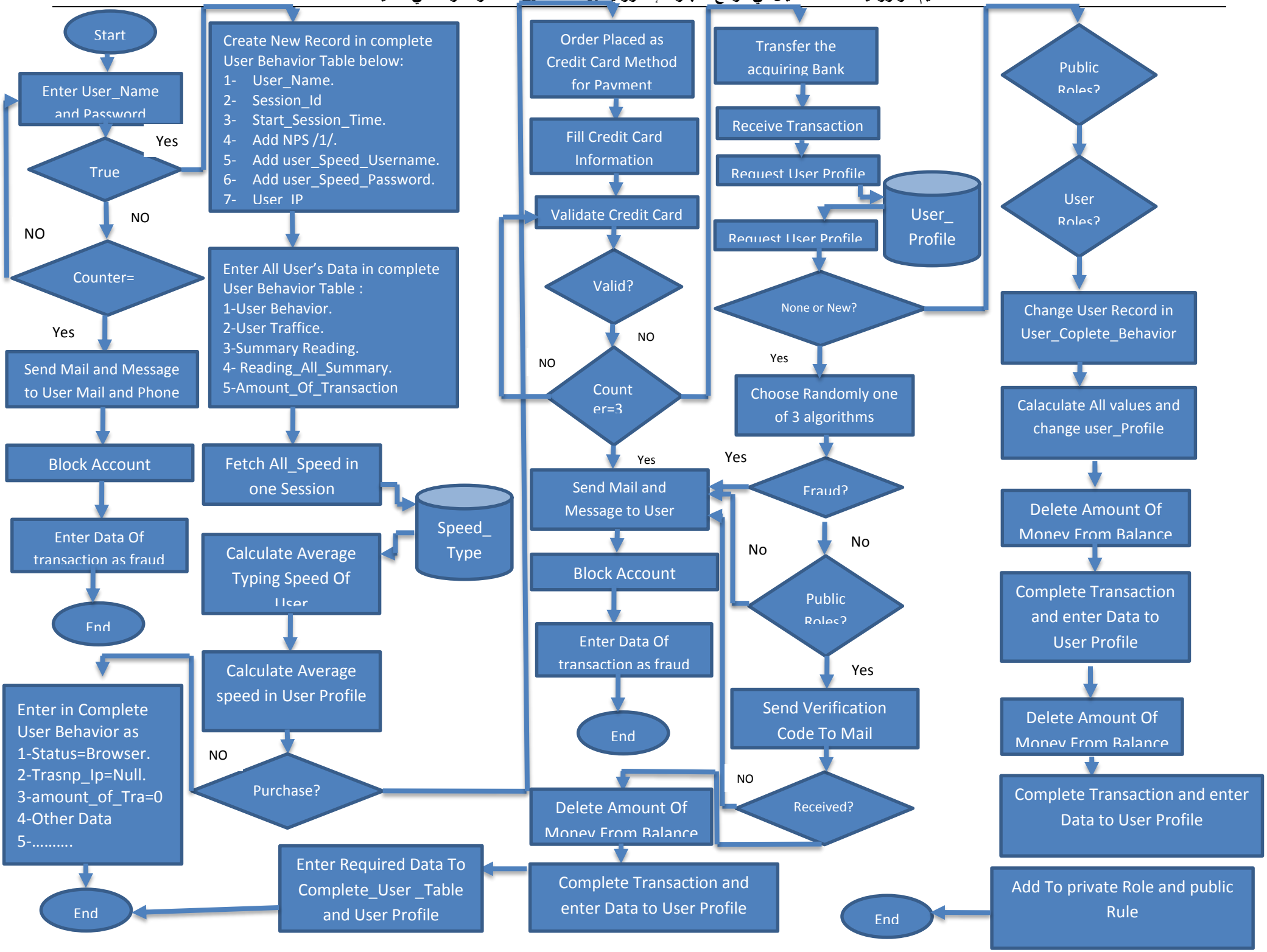
3- إعطاء بعض الخصائص أهمية أكثر من باقي الخصائص وهذه الخصائص هي:

- Speed_in_typing
- amount_in_transaction
- Average_Session_Time
- Direct_Purchase
- User_Type

وذلك من خلال تثقيل قيم هذه الخصائص بحيث يكون لها أثر أكبر من باقي الخصائص في تقييم المناقلة الحالية هل هي شرعية أم أنها عملية احتيالية.

نظراً لضرورة فهم آلية عمل النظام المقترح من قبل الباحث لا بد من شرح ألية عمله من خلال مخطط تدفقي للحل وإن كان المخطط كبير إلا أنه ضروري لفهم آلية عمل النظام المقترح:

تصميم خوارزمية كشف الاحتيال في مواقع التجارة الإلكترونية بهدف تحسين الدقة والسرعة في عمليات الكشف



قام الباحث باختبار النتائج التي قام بالحصول عليها بعد تطبيق النموذج المقترح على مجموعة بيانات kaggle

الأساسية والموسعة وذلك باستخدام كل من البرنامجين:

1- Weka: الذي يؤمن العديد من معايير الدقة التي يمكن الحصول عليها مباشرة من البرنامج، حيث يقوم البرنامج بأخذ عينات عشوائية من النتائج وتقييم الدقة بناءً على عدد حالات الاحتيالية المكتشفة، ويزودنا بقيم العديد من معايير الدقة المهمة.

2- Rstudio: والذي يعطي قيمة للدقة بشكل مباشر والتي تعبر عن عدد الحالات الاحتيالية المكتشفة الصحيحة .

ونظراً لكون مجموعة البيانات متوازنة ومتسقة وخالية من الضجيج بالتالي لم يكن من الضروري إدخال مجموعة البيانات في مرحلة معالجة البيانات ،وكانت النتائج على النحو التالي:

Correctly classified Instances	83725	97.9904%
Incorrectly Classified Instances	1717	2.0095469%
Kappa Statistic	0.7253	
Mean absolute error	0.00002	
Root mean squared error	0.0012	
Relative absolute error	11.9932%	
Root relative squared error	42.0012 %	
Total Number of Instances	85442	

===Detailed Accuracy By class ===

	TP	FP	Precision	Recall	F1	MCC	Roc	PRC	Class
	0.852	0.000	0.985	0.85	0.939	0.843	0.882	0.711	1
	1.000	0.112	1.000	1.000	1.000	0.843	0.882	1.000	0
AVG	0.979	0.235	0.979	0.999	0.979	0.843	0.8821	0.979	

الشكل(2) يوضح نتائج دقة الحل المقترح من قبل الباحث على قاعدة البيانات Kaggle باستخدام برنامج Weka

كما يظهر الشكل التالي نتائج تطبيق الحل المقترح على قاعدة البيانات بعد إضافة الخصائص المدروسة باستخدام برنامج RStudio وكانت النتائج على النحو التالي:

Confusion Matrix and Statistics

	Reference	
Prediction	Not Fraudulent	Fraudulent
Not Fraudulent	83695	1435
Fraudulent	0	31

الشكل(3) يوضح نتائج دقة الحل المقترح من قبل الباحث على قاعدة البيانات Kaggle باستخدام برنامج Rstudio

وبأخذ المتوسط الحسابي للدقة المقدمة من البرنامجين نلاحظ أن دقة الحل المقدم من قبل الباحث هي:

AVG Accuracy=(97.9904+97.955381)/2=97.97289%.

وبالتالي نلاحظ تحسن في الدقة من قبل الحل المقدم من قبل الباحث بفارق 0.516809% وهي دقة خوارزمية SVM، ولمزيد من الدقة قام الباحث بتطبيق الحل المقترح من قبله على قاعدة البيانات الموسعة واختبار النتائج على برنامجي WEKA,Rstudio وكانت النتائج على النحو التالي:

Correctly classified Instances	208842	98.5638 %
Incorrectly Classified Instances	2174	1.4361564%
Kappa Statistic	0.4125	
Mean absolute error	0.000252	
Root mean squared error	0.0092	
Relative absolute error	33.1263%	
Root relative squared error	46.0236 %	
Total Number of Instances	211885	

===Detailed Accuracy By class ===

	TP	TF	Precision	Recall	F−	MCC	Roc	PRC	Class
	0.786	0.000	0.953	0.686	0.999	0.804	0.912	0.911	1
	0932	0.533	0.999	0.953	0.999	0.804	0.912	0.994	0
AVG	0.982	0.522	0.988	0.953	0.999	0.804	0.912	0.995	

الشكل(4) يوضح نتائج دقة الحل المقترح من قبل الباحث على قاعدة البيانات Kaggle الموسعة باستخدام برنامج Weka

كما قام الباحث بتطبيق الحل المقترح على قاعدة البيانات الموسعة باستخدام برنامج Rstudio وكانت النتائج على النحو التالي:

Confusion Matrix and Statistics

	Reference	
Prediction	Not Fraudulent	Fraudulent
Not Fraudulent	208652	3002
Fraudulent	0	231

Accuracy : 98.4741732

الشكل(5)يوضح نتائج تطبيق الحل المقترح من قبل الباحث على قاعدة بياناتKaggle الموسعة

نلاحظ أن مقدار التحسن في الدقة عند تطبيق الحل المقترح من قبل الباحث على قاعدة البيانات الأساسية كان بمقدار 0. 516809% وفي المرة الثانية مقدار التحسن المتوسط للدقة بمقدار:

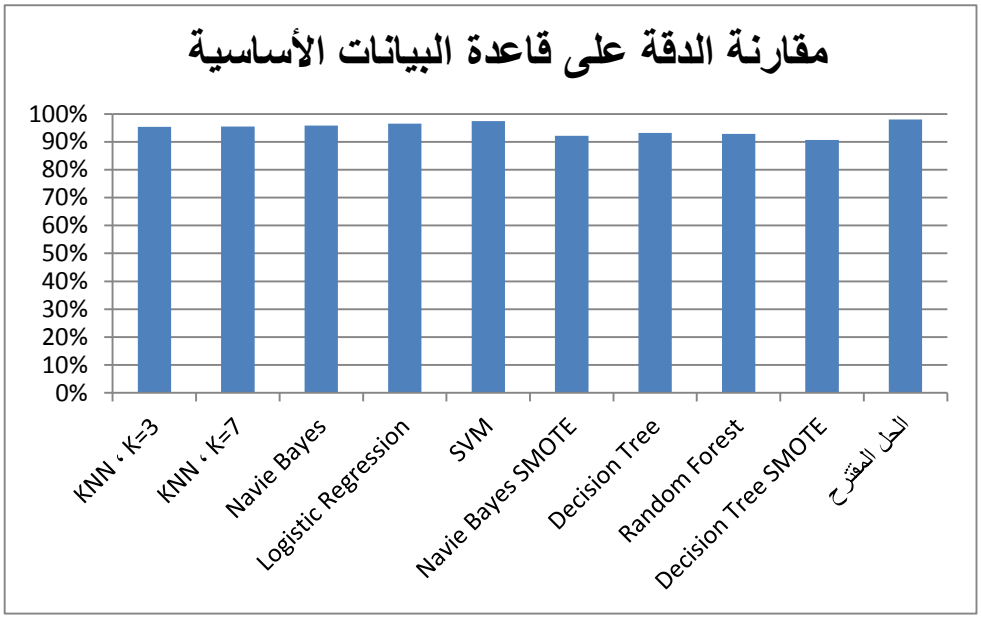
AVG Accuracy=(98.56938+98.4741732)/2=98.521776%

نلاحظ أن الدقة الناتجة عن الحل المقترح من قبل الباحث تتزايد بمرور الوقت والسبب هو زيادة عدد المعلومات الخاصة بالمستخدم ، حيث كلما زاد عدد مناقلات المستخدم مع الموقع كلما زادت الدقة لدينا نتيجة توفر معلومات أكثر عن المستخدم وبالتالي معرفة أكبر للنظام بسلوك المستخدم الحالي، وأن فرق الدقة بين الحل المقترح من قبل الباحث وبين أفضل خوارزمية يقدر بحوالي: 2.8819% وهذا مقدار كبير من ناحية الدقة، علماً بأن الحل المقترح من قبل الباحث ستتزايد دقته المقدمة بمرور الزمن.

مناقشة النتائج:

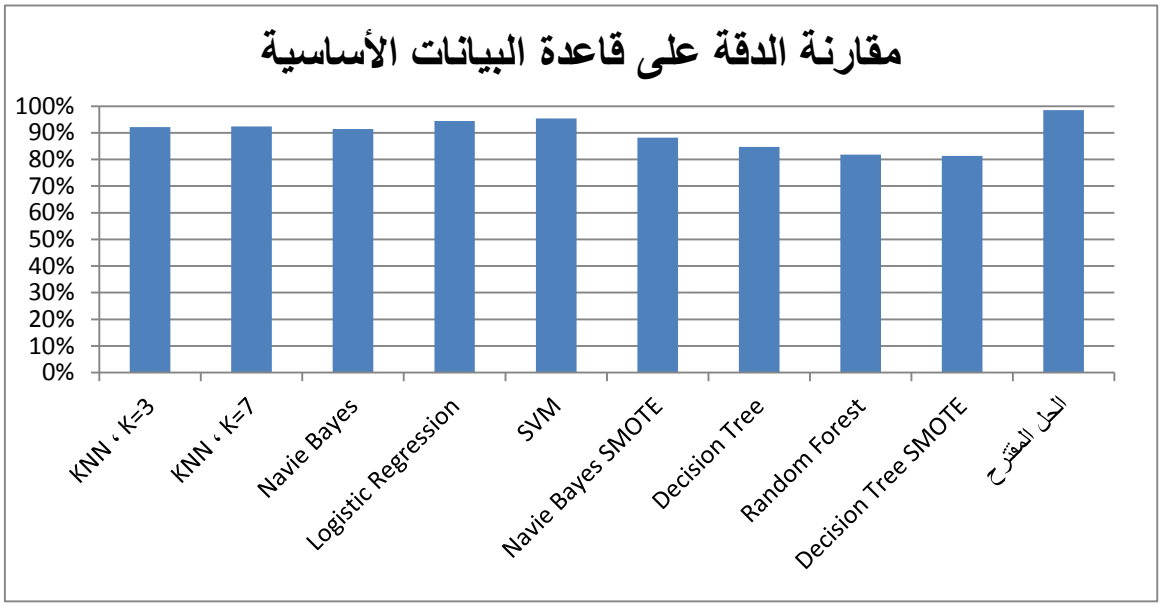
1- **الدقة:** تعتمد الدراسة والنظام المقدم من قبل الباحث على دراسة سلوك المستخدم وتفاعله مع الموقع لدينا، الأمر الذي يعطي النظام ميزة الدقة المتزايدة بمرور الوقت، حيث كلما ازدادت تفاعلات المستخدم مع النظام كلما تم الحصول على نتائج ذات دقة أعلى وكفاءة أفضل، حيث نلاحظ أن فرق الدقة بين الحل المقترح من قبل الباحث وأفضل خوارزمية كان بمقدار 0. 516809% عند تطبيق الحل على قاعدة البيانات الأساسية، وكان الفرق 2.8819% عند تطبيق الحل على قاعدة البيانات الموسعة، علماً بأن دقة الخوارزمية ازدادت بعد زيادة عدد سجلات قاعدة البيانات على عكس باقي الخوارزميات الأخرى والتي نقصت فيها الدقة بسبب عدم أخذ الخصائص المناسبة ضمن الحل، حيث كانت دقة الحل المقترح من قبل الباحث 97.97289% وأصبحت الدقة% 98.521776 بمقدار تحسن يبلغ 0.548886% .

يوضح الشكلان التاليان فرق الدقة بين الحل المقترح من قبل الباحث وبين الخوارزميات السابقة في مرحلتي المقارنة على قاعدة البيانات الأساسية وقاعدة بيانات **Kaggle:**



الشكل(6) يوضح مقارنة الدقة بين الحل المقترح والخوارزميات الأكثر استخداماً على قاعدة البيانات الأساسية

كما يظهر الشكل التالي مقارنة الدقة بين الحل المقترح من قبل الباحث والخوارزميات الأكثر استخداماً ،في حالة قاعدة البيانات الموسعة:



الشكل(7) يوضح مقارنة الدقة بين الحل المقترح والخوارزميات الأكثر استخداماً على قاعدة البيانات الموسعة

2- **السرعة:** تمت دراسة السرعة في هذا البحث من خلال ارتباط سرعة التنفيذ بعدد السجلات التي تقوم الخوارزميات بمسحها خلال كل عملية كشف للاحتيال، ومن المعروف أن عدد السجلات الموجودة في جدول المناقلات بين المستخدمين وموقع التجارة الإلكترونية أكبر بكثير من عدد سجلات جدول المستخدمين ضمن الموقع، ففي مجموعة البيانات الأساسية Kaggle كان عدد السجلات 284804 في جدول المناقلات بينما عدد السجلات الخاصة بالمستخدمين 4112 وفي مجموعة البيانات الموسعة عدد السجلات الخاصة بالمناقلات كانت ، وعليه فإن الحل المقدم من قبل الباحث يعتبر ذو سرعة أفضل من باقي

الخوارزميات من ناحية الزمن اللازم لمسح قاعدة البيانات حيث يقوم الحل المقترح من قبل الباحث على مسح جدول المستخدمين بينما تقوم باقي الخوارزميات بمسح كامل جدول المناقلات وذلك للكشف عن المناقلة الحالية هل هي شرعية أم أنها عملية احتيالية، وبالتالي فإن زمن التنفيذ الخاص بالحل المقترح من قبل الباحث أفضل بكثير من الخوارزميات السابقة.

لم يقم الباحث بدراسة تفصيلية للمقارنة بين زمن التنفيذ الخاص بالحل المقترح وبين زمن التنفيذ الخاص بالخوارزميات السابقة من الناحية العددية ، وعليه يجدر أن يكون هذا الأمر بحثاً مستقبلاً مهماً ولا سيما إذا أردنا التحسين على الخوارزميات السابقة أو على الحل المقترح من قبل الباحث.

3- **قابلية التطوير**: يتميز الحل المقدم من قبل الباحث بالاعتماد على مبدأ الاستقلالية في تقديم الحل ، حيث من الممكن التعديل على الحل بشكل كلي أو التطوير على جزء محدد دون الحاجة إلى إعادة بناء النظام ككل، حيث أن النظام يعتمد على جمع المعلومات عن المستخدم من خلال أجزاء برمجية مستقلة عن بعضها البعض وموزعة على كامل الموقع، وبالتالي فإن عملية التعديل على خاصية من الخصائص أو إضافة خاصية ما أو حذف خاصية يتطلب تعديل بسيط على قاعدة البيانات وتعديل على مكان واحد ضمن الموقع ، وبالتالي تتميز عملية التطوير بالسهولة والمرونة.

المراجع

1- PORWAL،U & MUKUND،S،2019- Credit card fraud detection in e-commerce. IEEE ،287P.

2- CAO، R & LIU، G& JIANG، C،2021- Two-level attention model of representation learning for fraud detection.IEEE، 1301P.

3- ZAHI،Y & SONG، W & LIU، X. &ZHAO، X،2018- A chi-square statistics based feature selection method in text classification. ICSESS،195P.

4- ALETH،K and SAMANTH،Y 2023-Using Kaggle.com database in fraud Detection System Using KNN and Navie Bayes. IEEE، P64.

5- PATIDAR،R & SHARMA،L،2021- Credit card fraud detection using neural network. International Journal of Soft Computing and Engineering (IJSCE)، 112P.

6- RAPTIDAR ,R, 2021- Fraud Detection using GA and AI.IEEE, P128.

7- KEVORT,A-2018. Faud Dtection Using Decision tree and Smote Decision Tree. IEEE , P83.

8- HOLLAND,J-2021. Using logistic Regression and KNN to detect fraud Transaction in e-Commerce. IEEE, P32.

9- ALI، S &ABD AL RAZAK،K &Othman، H، & Al-Dhaqm، M،2022- Financial fraud detection based on machine learning.Appl. Sci، 98P.

10- PATIDAR،R & SHARMA،L،2021- Credit card fraud detection using neural network. International Journal of Soft Computing and Engineering (IJSCE)، 112P.

11- LIUXGU، X، and SHANG،C، 2020-Quantitative detection of financial fraud based on deep learning with combination of E-commerce big data. IEEE P128.