اختبار أهم خوارزميات قواعد البيانات المعرفية المستخدمة في كشف الاحتيال وتحسين دقتها باستخدام قواعد البيانات المعتمدة على سلوك المستخدم

إعداد: علي إبراهيم

إشراف:د.بسيم عمران

ملخص

أنظمه الكشف عن الاحتيال: هي الأنظمة التي تهدف بشكل أساسي الى تحديد نوع المناقلة الحالية للمستخدم هل هي عمليه شرعيه أم عمليه نصب احتيالي، حيث تهدف أنظمة الكشف عن الاحتيال الى الحد من الآثار السلبية للاحتيال ودراسة طرق عملها وخوارزمياتها المتطورة بشكل دائم وذلك بهدف تقليل الخسائر المادية الناتجة عن عمليات النصب عن طريق الاحتيال.

قام الباحث بتحسين دقة خوارزميات قواعد البيانات المعرفية المستخدمة في كشف الاحتيال من خلال تطبيق هذه الخوارزميات على مجموعة بيانات معتمدة على سلوك المستخدم (من خلال دراسة خصائص المستخدم وسلوكه وتفاعله مع الموقع)، ومقارنة نتائج هذا الاختبار مع اختبار تطبيق هذه الخوارزميات على مجموعة بيانات لا تأخذ سلوك المستخدم كمعيار

أساسي في بنائها وكان هناك تحسن واضح في الدقة ولا سيما مع ازدياد عدد السجلات الموجودة في مجموعة البيانات.

الكلمات المفتاحية: قواعد البيانات المعرفية، كشف الاحتيال، قاعدة البيانات Kaggle، خوارزميات الكشف عن الاحتيال[1] .

Abstract:

Fraud detection systems: are systems which mainly aim to determine the type of user's current transaction, whether it is a legitimate or a fraud transaction. Fraud detection systems aim to reduce the negative effects of fraud and permanently study its working methods and advanced algorithms, to reduce its effects.

In this Article, researcher has improved the accuracy of KDD's algorithms by using user behavior's datasets, (this dataset has built by studying attributes, behavior and interactions of user with website), after that, he has compared the results of this study with results of applying KDD's algorithms on datasets which don't take user's behavior as basic standard in building. There was a clear improvement in accuracy, especially with the increase in the number of records in the database.

Key Words: KDD, Fraud Detection, Kaggle.com database, fraud detection algorithms

مقدمة

تعتبر دراسة الخصائص التي يتم عليها بناء مجموعة البيانات التي تتعامل معها خوارزميات نظام كشف الاحتيال من العوامل الأساسية والمهمة في نجاح عملية الكشف، وبالتالي فإن معرفة الخصائص التي تساعد في تحسين دقة هذه الأنظمة من أهم الأسس التي يجب أن تؤخذ بعين الاعتبار عند تصميم وبناء أنظمة كشف الاحتيال[2] .

قام الباحث بالاعتماد على منصة Kaggle العالمية التي تقدم طيف واسع من مجموعات البيانات التي لها علاقة بالاحتيال في مواقع التجارة الإلكترونية، وقام بتطبيق خوارزميات قواعد البيانات المعرفية على مجموعة البيانات هذه[3]، ودراسة تأثير زيادة عدد السجلات في انخفاض دقة النتائج، وقام الباحث بتحسين دقة هذه الخوارزميات وحل مشكلة انخفاض الدقة بازدياد عدد السجلات من خلال إضافة خصائص خاصة بسلوك المستخدم وتفاعله مع موقع التجارة الإلكترونية، ودراسة النتائج وتحليلها، فكان لها أثر إيجابي من ناحية تحسين الدقة بشكل جيد، وكانت طريقة ناجحة في حل مشكلة انخفاض الدقة بمرور الزمن وزيادة عدد السجلات.

مشكلة البحث

يعتبر التحسين على دقة النتائج الخاصة بنظام كشف الاحتيال الهدف الأساسي والمعيار الأهم في تقييم أداء أي نظام يستخدم لكشف الاحتيال في مواقع التجارة الإلكترونية، كما تعتبر بنية وطبيعة خصائص مجموعة البيانات التي يتعامل معها أي نظام لكشف الاحتيال من العوامل التي يجب أن تؤخذ بعين الاعتبار أثناء عملية بناء أنظمة كشف الاحتيال[4].

أهداف البحث

يهدف البحث إلى دراسة تأثير بنية وطبيعة البيانات التي يتعامل معها نظام كشف الاحتيال في دقة النتائج المقدمة، وذلك بهدف تحسين دقة النتائج ولا سيما مع ازدياد عدد السجلات ضمن مجموعة البيانات، من خلال إضافة خصائص المرتبطة بسلوك المستخدم على مجموعة البيانات وأثر هذه الخصائص في تحسين الدقة.

أهمية البحث

تأتي أهمية البحث من خلال دراسة أهم الخصائص التي تأخذ سلوك المستخدم بعين الاعتبار، وتبيان أثر بناء مجموعة بيانات (تعتمد على خصائص مرتبطة بسلوك المستخدم) في تحسين الدقة بشكل واضح، وجعل هذه الدقة تتزايد بمرور الزمن مع زيادة عدد السجلات المدروسة، الأمر الذي يعطي نظام لكشف الاحتيال يحقق النتائج المرجوة من بنائه ويخفف الخسائر المالية الكبيرة الناتجة عن عملية الاحتيال.

منهج وفرضيات البحث

يعتمد هذا البحث على المنهج التطبيقي وذلك من خلال استخدام برنامجي Weka, RStudio وتضمين الخوارزميات المدروسة وقاعدة البيانات ضمن البرنامجين ، ومقارنة الخوارزميات من ناحية معايير الدقة المطلوبة.

مجموعة البيانات Kaggle الأساسية والموسعة:

تقدم المنصة الإلكترونية Kaggle عدد كبير من مجموعات البيانات الخاصة بمواقع التجارة الإلكترونية كونها مفتوحة المصدر وتسمح بالحصول على مجموعات بيانات خاصة بمناقلات موقع تجارة إلكترونية وبشكل مجاني، وعليه قام الباحث بالاعتماد على مجموعة البيانات Credit Card Fraud Detection DataSet 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 الخوارزميات بتوسعة قاعدة البيانات وذلك من خلال زيادة عدد السجلات المدروسة إلى 532251 من خلال دمج مجموعة بيانات أخرى لموقع Kaggle بنفس الخصائص.

حيث قام الباحث بالاعتماد على مجموعة البيانات kaggle وذلك لكونها مجموعة بيانات غير 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	Product 2	transType

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

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 مجموعة البيانات

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 مجموعة البيانات

مجموعة البيانات المعتمدة على سلوك المستخدم:

اعتمد الباحث في بناء مجموعة البيانات المعتمدة على سلوك المستخدم على تسجيل تفاعلات المستخدم ضمن الموقع وتسجيل تحركاته ورغباته والمنتجات التي قام بالاطلاع

عليها أو التفاعل معها أو شرائها، حيث يوجد لدينا 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

هل المستخدم يقوم بتغيير IP بشكل مستمر أم لا	Changing_Ip
آخر عنوان IP للمستخدم	Last_IP
هل المستخدم يغير عنوان الشحن بشكل مستمر[6]	Changing_Tanspotation_location
هل المستخدم يقوم بالشراء مباشرة أم لا	Direct_Purchase
هل المستخدم يقوم بالشراء من أكثر من بطاقة ائتمانية	Purchase_from_many_Credit_Cards
هل المستخدم يطلب حسم أم لا	Required Discountig
الفترة الزمنية بين طلبي شراء(هل هنالك فترة زمنية بين طلبي شراء أم لا)	Times Between Transaction
هل المستخدم يقوم بشراء منتجات ذات أسعار مرتفعة	Purchase_Porduct
هل المستخدم يقوم بشراء منتجات متنوعة في نفس الجلسة أم لا	Product_Type
هل المناقلة الحالية شرعية أم غير شرعية	Trasaction Type

الجدول(4) 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 تم أخذ كل أصناف المستخدمين بعين الاعتبار وخاصة المستخدمين أصحاب الاهتمامات المتغيرة والسلوك المتغير بمرور الوقت، حيث يعتبر هذا النوع من أصعب أنواع المستخدمين من ناحية تقييم أداؤهم وسلوكهم، حيث يوجد لدينا العديد من أصناف المستخدمين[4] ، نذكر منها:

1- المستخدم الذي يتفاعل بشكل كبير خلال فترة زمنية قصيرة.
2- المستخدم الذي يتفاعل بشكل كبير خلال فترة زمنية طويلة.
3- المستخدم الذي يتفاعل بشكل قليل خلال فترة زمنية طويلة[5].
4- المستخدم الذي يتفاعل بسلوك ثابت وبأنواع محدد من المنتجات.
5- المستخدم الذي يتفاعل بسلوك متغير وبأنواع مختلفة من المنتجات[6].

حيث تم أخذ هذه الأصناف بعين الاعتبار خلال عملية المقارنة وبناء قاعدة البيانات الخاصة بسلوك المستخدمين.

مراحل الدراسة البحثية:

لقد تم تنفيذ الدراسة البحثية من خلال القيام بأربع مراحل من عمليات الاختبار التالية:

1- المرحلة الأولى: اختبار الخوارزميات على قاعدة البيانات الأساسية والتي تحتوي على 284808 سجل.
2- المرحلة الثانية : اختبار الخوارزميات على قاعدة البيانات الموسعة والتي تحتوي على 532251 سجل.
3- المرحلة الثالثة: اختبار الخوارزميات على قاعدة البيانات الخاصة بسلوك المستخدم والتي تحتوي على 320102 سجل.
4- المرحلة الرابعة: اختبار الخوارزميات على قاعدة البيانات الموسعة الخاصة بسلوك المستخدم والتي تحتوي على 634112 سجل.

أهم الدراسات المرجعية السابقة:

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

1- Fraud Detection using GA and AI -2021 ,RAPTIDAR,R [8] :
قام الباحث بالدمج بين الشـبكات العصـبونية والخوارزميات الجينية وقسـم الحل على ثلاث مراحل، حيث اعتمد في المرحلة الثانية على خوارزمية SVM لزيادة الدقة، وطبق الحل على قاعدة بيانات Kaggl، وعلى الرغم من الدقة التي تم الحصـــول عليها(95.923%) إلا أن النظام يتميز بالتعقيد وزمن التنفيذ العالي، إضـافة إلى الصـعوبة في التطوير والتعديل على الحل، حيث اعتمد النظام على تحديد المعلومات المستخدمة الخاصة ببطاقة الإئتمان من ناحية عنوان الشراء ، وقيمة الشـــراء ، وعنوان الشـــحن، ومقارنة المعلومات المذكورة مع المعلومات المخزنة ضـــمن قاعدة البيانات وذلك لاتخاذ قرار هل المناقلة شـــرعية أم أنها احتيالية، حيث نلاحظ أن المعلومات التي يتم التعامل معها في عملية الكشف لا تساعد بشكل كبير في عملية الكشف ، وبالتالي فإن إضافة خصائص لها علاقة أكبر بسلوك المستخدم سيكون له أثر واضح من ناحية تحسين الدقة في عملية كشف الاحتيال.

2- Keveort,A -2018. Faud Dtection Using Decision tree and Smote Decision Tree[9]قام الباحث بتقديم نظام يعتمد على خوارزمية شـــجرة دعم القرار وشـــجرة دعم القرار المعدلة وقام بتطبيق الحل على قاعدة البياناتkaggle، حيث قســـم الباحث الحل إلى مرحلتين : الأولى تعتمد على تطبيق خوارزمية شـجرة أخذ القرار المعدلة، ومن ثم تطبيق خوارزمية شـجرة دعم القرار وكانت دقة النتائج أقل من الحل الســـابق(94.281%)، وبزيادة عدد

السجلات سنحصل على دقة أقل وسيكون سرعة التنفيذ أكبر وبالتالي الحل غير مفيد في حال قواعد البيانات الضخمة.

3- يعتمد البحث Extracting and reasoning about implicit behavioral evidences for detecting fraudulent online transactions in e-Commerce للباحث Zhao, Jic [10] على دراسة السلوك المضمن للمستخدم ضمن موقع التجارة الإلكترونية واختبار النتائج من خلال *خوارزمية* Random Forest، حيث يعتمد الباحث على دراسة سلوك المستخدم من ناحية طبيعة المنتجات التي يتعامل معها، والقيمة الشرائية للمناقلة، والمعلومات الخاصة بعملية الشراء مثل عنوان IP التي تم الشراء منها ومقارنتها مع العنوان السابق، عنوان الشحن. تعتبر تلك المعلومات مهمة في عملية كشف الاحتيال، ولكن لم تأخذ الدراسة بعين الاعتبار إمكانية حصول المحتال على تلك المعلومات وخاصة معلومات بطاقة الائتمان من ناحية أن تكون بطاقة الائتمان مسروقة أو تم الحصول عليها من السوق السوداء، وإمكانية حصول المحتال على كمية كبيرة من المعلومات عن المستخدم من خلال التصيد الاحتيالي والهندسة الاجتماعية، بالتالي فإن الحصول على معلومات أكثر دقة عن تفاعل المستخدم مع الموقع سيعطي نتائج أفضل وسيكون له أثر كبير في زيادة دقة نظام كشف الاحتيال، كما أن استخدام خوارزميات مثل شجرة دعم القرار أو SVM سيكون له أثر أكثر إيجابية في دقة النتائج التي سيتم الحصول عليها من النظام المقدم من قبل الباحث.

4- كما تضمن البحث IDENTIFYING FRAUDLENT ACTIVITIES DETECTION IN E-COMMERCE WEBSITES للباحثين P V KUMAR , V. SAI GANESH, V. NAGARAJU and CH.VENKATESWARA RAO [11] دراسة لسلوك المستخدم ضمن الموقع

بالاعتماد على معلومات بطاقة الإئتمان، ومعلومات الجهاز الذي يقوم المستخدم بالشراء منه ، كما يضاف له وقت الشراء ، ووقت تسجيل الدخول وتطبيق هذه المعلومات على خوارزميات KNN,Decision Tree, Random Forest حيث أعطت الدراسة نتائج دقةٌ Random Forest=085,KNN =0.77,Decision Tree=0.77 ونلاحظ أن النتائج ما تحتاج إلى مزيد من التحسين سواء من ناحية المعلومات المدروسة أو الخوارزمية التي تتعامل مع هذه البيانات وذلك للحصول على نظام كشف يتميز بدقة متزايدة بزيادة عدد السجلات.

5- 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 [12] ، 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.

مما سبق يتبين أن أغلب الحلول المقدمة سابقاً لا تأخذ السلوك الفعلي للمستخدم مثل:

1- سرعة كتابة المستخدم.
2- تغيير عنوان IPالمستخدم بشكل دائم.
3- تغيير عنوان الشحن بشكل مستمر.
4- المدة الزمنية للجلسة.
5- قراءة المستخدم لمعلومات المنتج قبل شرائها .

وغيرها من المعلومات الأخرى بعين الاعتبار، وبالتالي فإن دراسة تأثر الدقة مع إضافة خصائص مرتبطة ارتباط وثيق بسلوك المستخدم سيكون موضوع البحث لدينا.

تستخدم خوارزميات قواعد البيانات المعرفية على نطاق واسع في مجال كشف الاحتيال، وفيما يلي أكثر هذه الخوارزميات استخداماً:[13] :

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

7- SVM.

8- Decision Tree.

9- Smote Decision Tree.

10- Logistic Regression.

11- Navie Bayes.

12- Smote Navie Bayes.

13- Random Forest.

وعليه سيقوم الباحث باختبار هذه الخوارزميات من حيث دقة الكشف وإجراء مقارنة فيما بينها بعد تطبيقها على كل من قاعدة البيانات الأساسية والموسعة ,ومن ثم تطبيق هذه الخوارزميات على قاعدة البيانات المعتمدة على سلوك المستخدم سواء الأساسية أو الموسعة وإعطاء النتائج.

المعايير المعتمدة لتحديد دقة الكشف:

من أشهر المعايير المستخدمة لتحديد الدقة في خوارزميات كشف الاحتيال في مواقع التجارة الإلكترونية [14] هي:

1- TP Rate.

2- FP Rate.

3- Precision= $TP/(TP + FP)$

ويعطى من خلال قسمة المناقلات التي تم تصنيفها على أنها مناقلات احتيالية وكان التقييم صحيح على مجموع المناقلات الاحتيالية صحيحة التقييم يضاف لها المناقلات التي تم تصنيفها أنها احتيالية وكان التقييم لها غير صحيح (مناقلة شرعية تم تصنيفها أنها احتيالية).

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

يعطى من خلال قسمة المناقلات الاحتيالية التي يتم تقييمها أنها احتيالية على مجموع المناقلات الاحتيالية صحيحة التقييم يضاف لها المناقلات الشرعية التي تم تقييمها أنها احتيالية.

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

حيث يعتبر معيار F1-Score من أهم المعايير التي ترتبط ارتباط وثيق بدقة النتائج المقدمة، حيث يعتمد على دراسة كل الحالات الخاصة بالدارسة سواء من ناحية الحالات الهدف(المناقلات الاحتيالية)، أو الحالات التي ليست الهدف(الحالات الشرعية).

لقد تم استخدام برنامجي Weka,Rstudio[15] من أجل تنفيذ الدارسة البحثية واختبار الخوارزميات الثماني في المراحل الأربع المذكورة سابقاً.وتم الحصول على النتائج التالية:

المرحلة الأولى: بعد تطبيق الخوارزميات السابقة على برنامجي WEKA,RStudio تم الحصول على نتائج الاختبار. حيث يوضح الشكلان 1-2 نتائج التطبيق على البرنامجين لخوارزمية الجار الأقرب عند قيمة K=3 . حيث سيتم الاكتفاء بوضع النتائج ضمن جداول وليس ضمن واجهات لاختصار عدد صفحات البحث.

Correctly classified Instances		81540		95.433%					
Incorrectly Classified Instances		3902		4.567%					
Kappa Statistic				0.833					
Mean absolute error				0.0008					
Root mean squared error				0.024					
Relative absolute error				21.9704%					
Root relative squared error		53.2988%							
Total Number of Instances		85442							
===Detailed Accuracy By class ===									
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	Roc	PRC	Class
	0.747	0.000	0.942	0.747	0.833	0.839	0.911	0.779	1
	1.000	0.253	0.999	1.000	1.000	0.839	0.911	1.000	0
AVG	0.999	0.252	0.999	0.999	0.999	0839	0.911	0.999	

الشكل(1)يوضح نتيجة تطبيق خوارزمية KNN بقيمة K=3 على قاعدة البيانات الأساسية باستخدام برنامج Weka

وقام الباحث بتطبيق خوارزمية الجار الأقرب على قاعدة البيانات الأساسية باستخدام برنامج Rstudio وكانت النتائج على النحو التالي:

Confusion Matrix and Statistics		
	Reference	
Prediction	Not Fraudulent	Fraudulent
Not Fraudulent	81449	2399
Fraudulent	0	1594
Accuracy : 95.323%		

الشكل(2) **يوضخ نتيجة تطبيق خوارزمية KNN بقيمة K=3 على قاعدة البيانات الأساسية باستخدام برنامج RStudio**

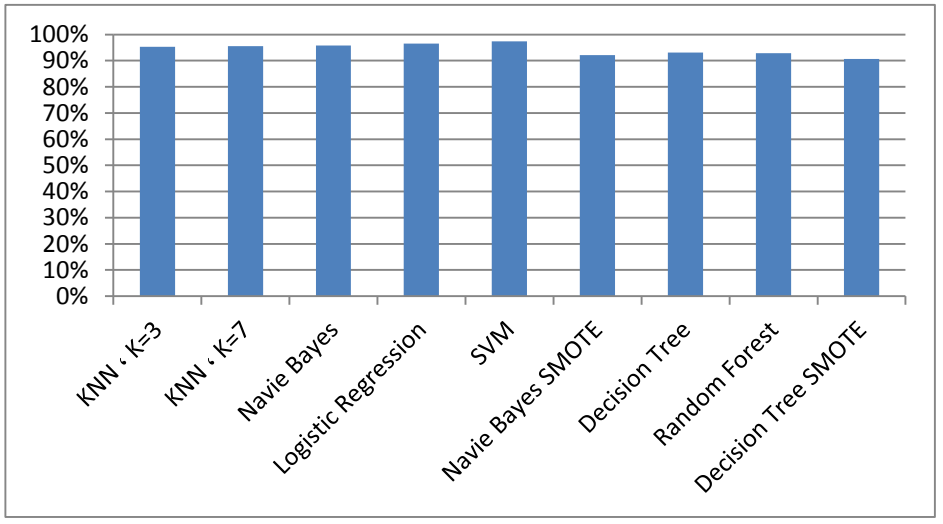
بعد إجراء الاختبار على جميع الخوارزميات السابقة باستخدام برنامجي Weka,RStudio تم الحصول على النتائج المبينة في الجدول(5) التالي:

	Accuracy	AVG

Algorithms	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(5): نتائج اختبار الخوارزميات على قاعدة البيانات الأساسية

كما يبين الشكل(3) مخططاً بيانياً لنتائج اختبار الخوارزميات السابقة على قاعدة البيانات الأساسية:



الشكل(3) مخطط بياني لنتائج اختبار الخوارزميات السابقة على قاعدة البيانات الأساسية

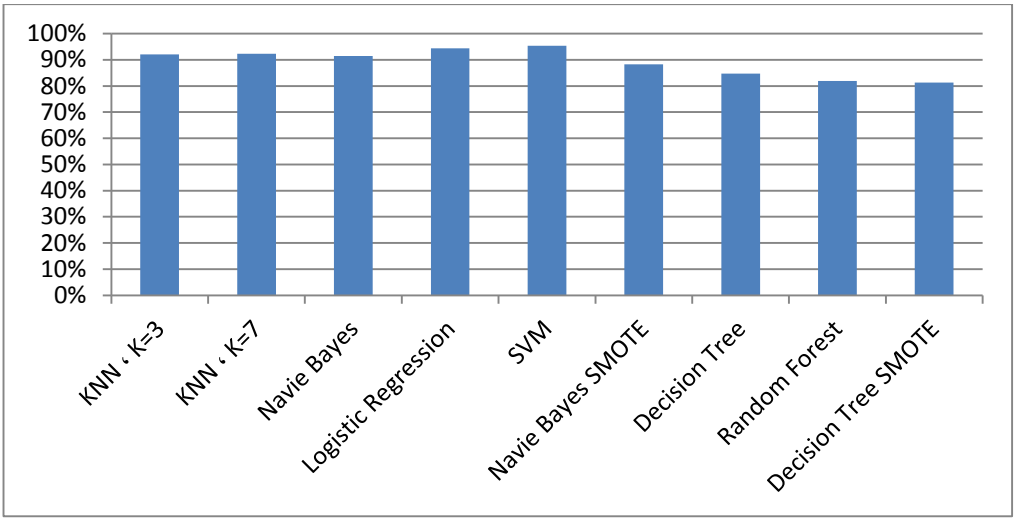
مما سبق يتبين أن أفضل الخوارزميات من ناحية الدقة هي خوارزمية SVM حيث بلغت دقتها 97.456% وأن أقل الخوارزميات دقة هي خوارزمية Decision Tree Smote بمعدل دقة 90.678235% وبالتالي تعتبر خوارزمية SVM هي أفضل خوارزمية من الخوارزميات المستخدمة في كشف الاحتيال.

المرحلة الثانية: في هذه المرحلة من الدراسة البحثية، اقترح الباحث زيادة عدد السجلات الخاصة بقاعدة البيانات من خلال إضافة قاعدة بيانات أخرى من Kaggle إلى قاعدة البيانات الأساسية بحيث أصبح عدد السجلات الكلي 532251 وقام باختبار الخوارزميات الثماني المذكورة سابقاً على قاعدة البيانات الموسعة وبعد تطبيقها على برنامجي Weka,Rstudio تم الحصول على النتائج المبينة في الجدول(6):

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

كما يبين الشكل(4) مخططاً بيانياً لنتائج اختبار الخوارزميات السابقة على قاعدة البيانات الموسعة:



الشكل(4) مخطط بياني لنتائج اختبار الخوارزميات السابقة على قاعدة البيانات الموسعة

مما سبق يتبين أن دقة جميع الخوارزميات المدروسة قد انخفضت وبقيت خوارزمية SVM هي الخوارزمية الأفضل من حيث الدقة.

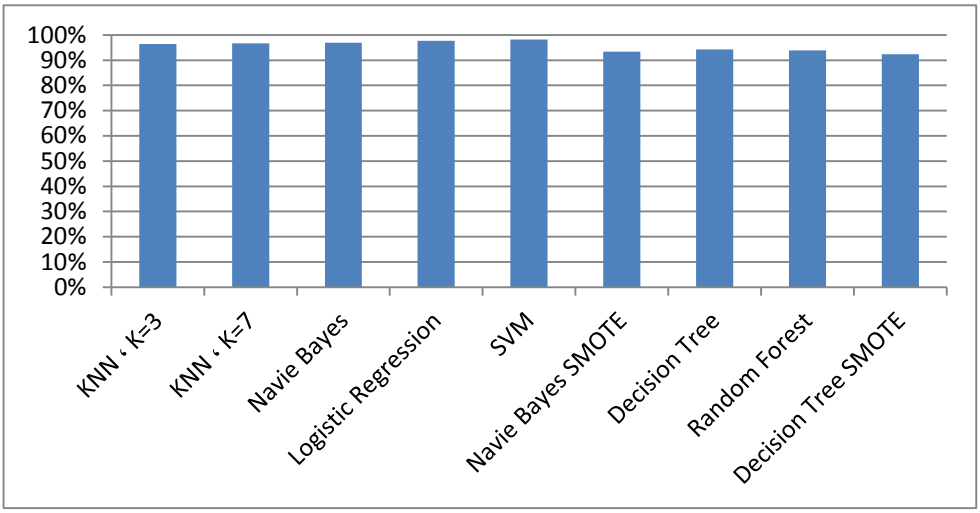
إن سبب هذا الانخفاض يعلل من خلال عدم كفاية الخصائص المدروسة, وعد أخذ الخصائص التي ترتبط ارتباط وثيق الحالة المدروسة بعين الاعتبار، وهذا ما دفع الباحث أخذ الخصائص التي لها علاقة بسلوك المستخدم وتفاعله مع الموقع لدنيا بعين الاعتبار والهدف من ذلك دراسة دقة هذه الخوارزميات في حال تغيير نمط ونوع البيانات التي تتعامل معها ودارسة أثر هذه الإضافة على دقة الخوارزميات بمرور الزمن مع ازدياد عدد السجلات الخاصة بمناقلات المستخدم، وبالتالي قام الباحث بالاعتماد على قواعد البيانات التي تعتمد بشكل أساسي على جمع وتخزين أكبر كمية من المعلومات عن المستخدم وهذا ما قام به في المرحلتين الثالثة والرابعة.

المرحلة الثالثة: قام الباحث بجمع المعلومات عن المستخدمين وذلك من خلال تسجيل كل تفاعلات المستخدم وتحركاته والمنتجات التي تفاعل معها سواء من ناحية قراءة ملخص المنتج أو شرائه أو البحث عن منتجات مشابهة للمنتج الحالي، كل هذه المعلومات تم تخزينها ضمن قاعدة البيانات وتم الحصول على النتائج المبينة في الجدول(7) بعد اختبارها على برنامجيً Weka,Rstudio :

	Accuracy		AVG
Algorithms	Weka	Rstudio	
K=3،KNN	96.562%	96.421 %	96.4915%
K=7،KNN	96.722%	96.689%	96.7055 %
Navie Bayes	96.988%	96.896%	96.942%
Logistic Regression	97.742%	97.682%	97.712%
SVM	98.212%	98.102%	98.157%
Navie SMOTE	93.454%	93.3712%	93.4126%
Decision Tree	94.252%	94.201%	94.2265%
Random Forest	93.992%	93.912%	93.952%
Decision Tree SMOTE	92.431%	92.391%	92.411%

الجدول(7): نتائج اختبار الخوارزميات على قاعدة البيانات المعتمدة على سلوك المستخدم

كما يبين الشكل(5) مخططاً بيانياً لنتائج اختبار الخوارزميات السابقة على قاعدة البيانات المعتمدة على سلوك المستخدم:



الشكل(5) مخطط بياني لنتائج اختبار الخوارزميات السابقة على قاعدة البيانات المعتمدة على دراسة سلوك المستخدم

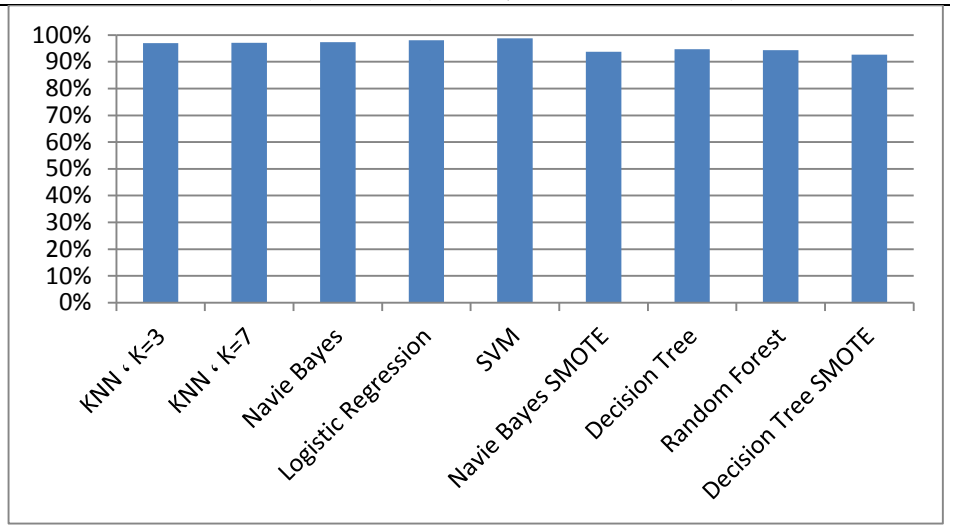
يلاحظ من نتائج الدراسة في المرحلة الثالثة تحسن في الدقة نتيجة تعامل الخوارزميات مع قاعدة بيانات خاصة بسلوك المستخدمين، هذا التحسن سيكون له دور كبير في التحسين على عمل خوارزميات قواعد البيانات المعرفية، نلاحظ هنا أن ازدياد المعرفة الخاصة بسلوك المستخدم وتفاعله مع قاعدة البيانات من خلال إضافة خصائص أكثر ترتبط ارتباط وثيق بسلوك المستخدم بالإضافة إلى المعلومات الأساسية الخاصة ببطاقات الائتمان السبب الأساسي في ازدياد الدقة المدروسة.، ومن المهم دراسة أثر هذه الإضافة على دقة النتائج بمرور الزمن ومع ازدياد عدد السجلات الخاصة بسلوك المستخدم، وهذا ما قام به الباحث في المرحلة الرابعة.

المرحلة الرابعة: في هذه المرحلة قام الباحث باختبار الخوارزميات الثماني السابقة على قاعدة البيانات المعتمدة على سلوك المستخدم وذلك بعد زيادة عدد السجلات واختبار النتائج على برنامجي Weka,Rstudio وكانت النتائج كما هو مبين في الجدول(8):

	Accuracy		AVG
Algorithms	Weka	Rstudio	
K=3،KNN	96.964 %	96.932 %	96.948%
K=7،KNN	97.102%	97.098%	97.1%
Navie Bayes	97.343%	97.251%	97.297%
Logistic Regression	98.112%	98.073%	98.0925%
SVM	98.802%	98.779%	98.7905%
Navie SMOTE	93.763%	93.6441%	93.70355%
Decision Tree	94.6905%	94.6211%	94.6558%
Random Forest	94.378%	94.312%	94.345%
Decision Tree SMOTE	92.7122%	92.7012%	92.7067%

الجدول(8): نتائج اختبار الخوارزميات على قاعدة البيانات المعتمدة على سلوك المستخدم مع زيادة عدد السجلات

كما يبين الشكل(6) مخططاً بيانياً لنتائج اختبار الخوارزميات السابقة على المعتمدة على سلوك المستخدم مع زيادة عدد السجلات :



الشكل(6) مخطط بياني لنتائج اختبار الخوارزميات السابقة على قاعدة البيانات المعتمدة على سلوك المستخدم مع زيادة عدد السجلات

يلاحظ من الدراسة والاختبار في المرحلة الرابعة أن الدقة ازدادت بزيادة عدد السجلات ، حيث ازدادت دقة خوارزمية SVM بمقدار 0.6335% والسبب في ذلك هو كون الخوارزميات تتعامل مع قاعدة بيانات ديناميكية مرتبطة بشكل أساسي بسلوك المستخدم، حيث تقدم قاعدة البيانات هذه معلومات أكثر دقة عن المستخدمين ، الأمر الذي يسبب زيادة في دقة هذه الخوارزميات بزيادة عدد السجلات وذلك نتيجة ازدياد المعلومات الخاصة بالمستخدمين.

مناقشة النتائج:

من خلال نتائج الدقة التي تم الحصول عليها في مراحل الاختبار الأربعة يمكن استنتاج ما يلي:

1- تلعب طبيعة البيانات التي تتعامل معها خوارزميات الكشف عن الاحتيال دور مهم وأساسي في دقة النتائج المقدمة وفي ازدياد هذه الدقة مع ازدياد عدد

المناقلات(السجلات)الخاصة بالمستخدم، أي أن الدقة تزداد كلما زادت الكمية المعرفية عن سلوك المستخدم.

2- يجب أخذ السلوك الفعلي للمستخدم ضمن الخصائص المدروسة في تقييم نوعية المناقلة الحالية هل هي شرعية أم أنها احتيالية، بحيث نجعل سلوك المستخدم المعيار الأساسي في عمليات التقييم، ودراسة سلوك المستخدم يجب أن تتم على كامل المناقلات التي صنفت على أنها شرعية ، ويجب أخذها بعين الاعتبار بشكل كامل في عمليات التقييم.

3- إن SVM هي الأفضل من بين جميع الخوارزميات الثماني الأكثر استخداماً، وكان لتفاعل هذه الخوارزمية مع مجموعة بيانات خاصة بسلوك المستخدم دور كبير وإيجابي في تحسن دقة كامل الخوارزميات المدروسة ولا سيما خوارزمية SVM

4- إن اعتماد خوارزميات قواعد البيانات المعرفية على قواعد البيانات التقليدية لا يعطي النتائج المرجوة من ناحية الدقة وخاصة مع ازدياد عدد السجلات.

5- من الجدير بالذكر أن التعامل مع قواعد البيانات الديناميكية يسبب زيادة في حجم التخزين أكثر من قواعد البيانات التقليدية الأمر الذي يتطلب حل لحجم التخزين المتزايد، عادة ما يتم الاعتماد على التخزين السحابي للتعامل مع مثل هذه القواعد.

المراجع

1- CARNEIOR,N, & FIGUEIREA,G and M. Costa,2018- A data mining based system for credit-card fraud detection .Decis. Support Syst P132.

2- RESHMA,F & SIFATULLAH,S ,2022 Credit Card Fraud Detection Using Data Mining . IJCRT.P23.

3- https://www.kaggle.com/datasets/ranjeetshrivastav/fraud-detection-dataset.

4- MEKHAK,M & SANDEEP ,S 2019 Detect Frauds in Credit Card using Data Mining Techniques. IJITEE, P121.

5- Y. FESTA,Y & VOROBEYEV,I 2022- A hybrid machine learning framework for e-commerce fraud detection, Model Assist. Stat. Appl,P75.

6- KHAN, A, & B. MISHRA,F ,2022-Developing a credit card fraud detection model using machine learning approaches. Int. J. Adv. Comput. Sci. Appl P418.

7- M. ZAMINI,Z and G. MONTAZER,G,2018- Credit card fraud detection using autoencoder based clustering , SVM and Logistic Regression. Int. Symp. Telecommunications, Tehran, Iran, P28.

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

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

10- Zhao, Jic, et al. "Extracting and reasoning about implicit behavioral evidences for detecting fraudulent online transactions in e-Commerce. SCiOpen,2024

11- P V KUMAR , V. SAI GANESH, V. NAGARAJU and CH.VENKATESWARA RAO, IDENTIFYING FRAUDLENT ACTIVITIES DETECTION IN E-COMMERCE WEBSITES, JETIR,2024,V11.

12- Patrik,G and Hugo,S, 2024 Detecting Fraudulent User Behavior UPTEC p55.

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

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

15- DORN,S and S. GEETHA,S,2020- Credit card fraud detection using machine learning algorithms. Procedia Computer Science, p45.