Use of Data Mining Techniques to Detect Fraud in Procurement Sector

استخدام تقنيات التنقيب عن البيانات للكشف عن الاحتيال في قطاع المشتريات

by

SUMAYYA ABDULLA AL HAMMADI

Dissertation submitted in fulfilment

of the requirements for the degree of

MSc INFORMATICS

at

The British University in Dubai

January 2022

DECLARATION

I warrant that the content of this research is the direct result of my own work and that any use made in it of published or unpublished copyright material falls within the limits permitted by international copyright conventions.

I understand that a copy of my research will be deposited in the University Library for permanent retention.

I hereby agree that the material mentioned above for which I am author and copyright holder may be copied and distributed by The British University in Dubai for the purposes of research, private study or education and that The British University in Dubai may recover from purchasers the costs incurred in such copying and distribution, where appropriate.

I understand that The British University in Dubai may make a digital copy available in the institutional repository.

I understand that I may apply to the University to retain the right to withhold or to restrict access to my thesis for a period which shall not normally exceed four calendar years from the congregation at which the degree is conferred, the length of the period to be specified in the application, together with the precise reasons for making that application

Signature of the student

COPYRIGHT AND INFORMATION TO USERS

ABSTRACT

Procurement is an extensive and complex sector in the manufacturing industry, and has attracted an extensive and wide-spreading fraud that directly impacts the operation of an organization and economy at large. These fraudulent activities have contributed to rising problems in the manufacturing industry. Several fraud detection systems are being used in the procurement and logistics sector, and their challenge is incapability of realizing the burden of the money lost and abnormal behaviors in the procurement process. Another major problem with the current system is that having an ever-growing amount of data requires a proportional growing number of staff members to analyze the data. In addition, some of the organizations carry out this task manually using their specialized staff.

Despite the implementation of various strategies aiming to fight and reduce fraud in the procurement sector, such as random and periodic audits, whistle-blowing, and many others, most of the UAE's organization still uses the manual approach to do these audits and monitoring of the procurement process. This has continued to be a challenge in most of the businesses in the UAE.

This research aims to analyze the reliability and efficiency of data mining techniques in detecting and preventing fraud in the procurement sector in the UAE and globally. The method used in this research is a classification of models and algorithms used in data mining. All techniques also will be studied; they include clustering, tracking patterns, classifications and outlier detection.

From this study I found out that most of the organizations lose quite a huge amount through fraud in their procurement sector. However, unsupervised data mining techniques are reliable in

detecting fraud before they happen. For the research, I found out the importance of data mining in detecting fraud in procurement. Data analytics reflects on the structuring of the data to be usable and accessible to teams or individuals who require information about procurement in a company. This essentially makes it easy to detect fraud and thus prevents it from happening.

The findings from this study will help implement a system that will significantly reduce fraud in the procurement sector. It will save companies a lot of money which will positively impact. This study concluded that most of the companies lose money due to fraud. They are willing to invest their money in fraud detection and control systems that will curb fraud. Fraud detection is a field that requires dynamic research and periodical upgrades and innovations because fraudsters are many and skilled; they consistently devise new ways to perform fraud in a less detectable way. From this study, I found out that the use of data mining techniques will help discover entirely invisible patterns and alert the fraudsters. There is a need for the companies to acquire new technology devices and ways to mitigate fraud in their procurement sector.

Keywords: data mining, fraud, techniques, procurement, prevention, protection, classification, pattern, affect, detection.

قطاع المشتريات هو قطاع واسع وممتد ومعقد في مجال الصناعة التحويلية، وقد واجهت المشتريات عمليات احتيال واسعة النطاق والانتشار مما أثرت سلباً بشكل مباشر على عمل المؤسسات والاقتصاد بشكل عام. وقد ساهمت هذه الأنشطة الاحتيالية في زيادة التحديات والصعوبات والثغرات في الصناعة التحويلية وقد تم استخدام العديد من أنظمة الكشف عن الاحتيال في قطاع المشتريات والخدمات اللوجستية ، ويتمثل التحدي في عدم القدرة على إدراك عبء الأموال المفقودة والسلوكيات غير الطبيعية في عملية الشراء مما سببت مشكلة رئيسية أخرى في النظام الحالي هي أن وجود كمية متزايدة من قواعد البيانات يتطلب عددًا متزايدًا نسبيًا من الموظفين لتحليل وتقييم البيانات. بالإضافة إلى ذلك ، تقوم بعض المؤسسات بهذه المهمة يدويًا ولكن باستخدام موظفيها المتخصصين والمعنيين بذلك.

على الرغم من تنفيذ الاستراتيجيات المختلفة التي تهدف إلى مكافحة الاحتيال والحد منه في قطاع المشتريات من مثل عمليات التدقيق العشوائية والدورية، والإبلاغ عن المخالفات، وغيرها الكثير ، إلا أن معظم المؤسسات في دولة الإمارات العربية المتحدة لا تزال تستخدم النهج اليدوي للقيام بعمليات التدقيق هذه والمراقبة على عمليات الشراء وإلى الآن ظل هذا يمثل تحديًا في معظم الشركات في الإمارات العربية المتحدة.

يهدف هذا البحث إلى تحليل موثوقية وكفاءة تقنيات التنقيب عن البيانات في اكتشاف ومنع الاحتيال في قطاع المشتريات في دولة الإمارات العربية المتحدة وعلى مستوى العالم. ودراسة الطريقة المستخدمة في هذا البحث هي تصنيف النماذج والخوارزميات المستخدمة في التنقيب عن البيانات وسيتم أيضًا دراسة جميع التقنيات؛ وهي تشمل التجميع وأنماط التتبع والتصنيفات والكشف عن الحالات الغير مألوفة.

من هذه الدراسة تم الوصول إلى حقيقة بارزة وهي أن معظم المؤسسات تخسر قدرًا كبيرًا جدًا من خلال الاحتيال في قطاع المشتريات لديها. ومع ذلك، فإن تقنيات التنقيب عن البيانات غير الخاضعة للرقابة يمكن الاعتماد عليها في اكتشاف الاحتيال قبل حدوثه. وعلاوة غلى ذلك، اكتشفنا أهمية التنقيب عن البيانات في اكتشاف الاحتيال في المشتريات حيث أنها تنعكس على تحليلات البيانات ومدى أهميتها من خلال العمل على هيكلة البيانات لتكون قابلة للاستخدام ويمكن الوصول إليها من قبل الفرق أو الأفراد الذين يحتاجون إلى معلومات حول المشتريات في الشركة مما يجعل من السهل اكتشاف الاحتيال وبالتالي يمنع حدوثه ومواجهته قبل وقوعه.

ستساعد نتائج هذه الدراسة في تنفيذ نظام من شأنه أن يقلل بشكل كبير من الاحتيال في قطاع المشتريات. سيوفر على الشركات الكثير من المال مما سيؤثر بشكل إيجابي. تلخّصت هذه الدراسة إلى أن معظم الشركات تخسر المال بسبب عمليات الاحتيال وأن عليهم أن يكونوا على استعداد لاستثمار أموالهم في أنظمة الكشف عن الاحتيال والتحكم فيه التي من شأنها تمنع مواجهتها. كشف الاحتيال هو مجال يتطلب بحثًا ديناميكيًا وترقيات وابتكارات تكنولوجية دورية لأن المحتالين كثيرون وماهرون؛ يبتكرون باستمرار طرقًا جديدة للقيام بالاحتيال بطريقة أقل قابلية للاكتشاف. من هذه الدراسة، نؤكّد على أن استخدام تقنيات التنقيب عن البيانات سيساعد في اكتشاف أنماط غير مرئية تمامًا وتنبيه المحتالين و هناك حاجة للشركات للحصول على أجهزة تقنية جديدة وطرق للتخفيف من الاحتيال في قطاع المشتريات الخاصة بهم.

Table of Contents

DECLARATION 1

ABSTRACT 4

CHAPTER ONE: INTRODUCTION 1

Introduction 1

Background of study 1

Problem Statement 2

Objectives of the Study 2

Main Objective 2

Specific Objective 2

Research question 3

Significance of the study 3

Scope 3

Organization of the paper 3

CHAPTER TWO: LITERATURE REVIEW 5

Introduction 5

Fraud Detection Literature by Subject 7

Procurement Fraud and Detection 7

Procurement Fraud Detection and Prevention 9

Credit Card Fraud and Detection 11

Medical and Insurance Fraud and Detection. 12

Data Mining Techniques in Fraud Detection 12

Tracking patterns 13

Classification 13

Outlier detection 13

Clustering 13

Reasons and applications of clustering in data mining 14

Association Components 15

Association rules, algorithms, and the use in fraud detection 15

Supervised Data Mining Methods 16
Unsupervised Data Mining Methods 16
Data Mining Classification Techniques 19
Naïve Bayes 19
Decision Tree 20
Research Studies Data Mining in Fraud Detection 20
Summary 21
CHAPTER THREE: METHODOLOGY 24
Introduction 24
Dataset 24
Algorithm: 26
K-modes Clustering 26
Algorithm: 26
Clustering on a Hierarchical Scale 27
Algorithm: 27
Medoids-Related Partitioning 28
Algorithm: 28
Self-Organizing Maps (SOM). 29
Algorithm: 29
Research Design 30
Research Timeline 30
Business Understanding 30
Data analytical tool and technique 30
Research sampling 30
Evaluation 32
Deployment 32
Ethical consideration 32
Methodology limitations and challenges 32
CHAPTER FOUR: PRIMARY AND SECONDARY DATA 34
Introduction 34

Secondary Sources 35

CHAPTER FIVE: DATA ANALYSIS 36

Implementations 36

Techniques for Clustering. 37

Clustering via K-means: 37

Clustering Via K-Mode: 38

Clustering via Medoids Partitioning 40

Clustering via SOM: 41

Unsupervised Data Mining Methods 42

Similarity in DISC. 42

SRA Performance Evaluation. 43

Detection of Fraud Automatically. 43

Additional Validation of the SRA Ranking. 43

CHAPTER SIX: RESULTS AND DISCUSSION 46

Data Mining in Procurement Fraud Detection 46

Data Mining Techniques Used in Analysis and Detection of Procurement Fraud 47

Fraud Activities in the Procurement Sector 47

CHAPTER SEVEN: CONCLUSION AND RECOMMENDATION 49

Recommendation 49

Summary of the Study 49

Recommendation of Future Work 50

Conclusion 50

REFERENCES 52

CHAPTER ONE: INTRODUCTION

Introduction

This chapter presents the background study, the statement of problems, objectives of the research, the research questions, scope and significance of this study.

Background of study

Many organizations strive to serve their customers with the most efficient and value-oriented services in whichever industry they operate. This involves streamlining the different departments in an organization to deliver on the best possible outcome that, in the end, sells the reputation of the company or the organization to the consumers, who are the ultimate judges of the products and services. Even though management is generally in the spotlight when it comes to the organization's general performance, each department significantly contributes to the collective output of the organization. Procurement of goods in an organization, for example, weighs about 60 per cent of the organization's revenue, varying from the different sectors of the economy (Al Hosani, 2020).

It is, therefore, prudent for companies to invest in rooting out inadequacies in their procurement by optimizing the operation rules in their procurement departments. Without a keen eye on the procurement in an organization, savings will not only remain unconsolidated but the room is created for harmful practices that are, in most cases, intentional and are the ones termed as procurement fraud The fraud takes many fronts in and out of the organization. The most common is the collaboration between the insiders, the requesters and the external entities who are the vendors. Both of these parties take advantage of the unnoticed agreements and arrangements to do procurement that, in most cases, create monetary losses within an organization (Carnairo,2020).

Problem Statement

Transparency is a vital tool when it comes to fraud elimination and it is essential in an organization's procurement process. Anomalies detection in procurement activities is essential to find, prevent, and take actions of possibly misappropriated funds. Considering the procurement department as a potential source of misuse of the organization's funds, performing regulatory control to ensure that the procurement processes are in line with regulatory procedures is an essential task. Currently, the organization carries out this task manually using its specialized staff. The main problem with the current system is that having an ever-growing amount of data requires a proportional growing number of staff members to analyze the data (Hardinata,2021).

Objectives of the Study

Main Objective

The use of data analysis using data mining techniques, as unsupervised learning, for anomaly detection in procurement activities will serve as an automated tool for implementing procurement fraud detection processes that need in-depth verification. This research addressed this problem by proposing the automation of control tasks for fraud detection in the procurement process by comparing two unsupervised data mining techniques.

Specific Objective

1. To identify fraud activities in the procurement sector
2. To study data mining techniques that can be implemented to curb fraud in the procurement sector.
3. To propose a reliable automation control system to be used in detecting fraud in the procurement sector.

Research question

i. How effective is association compared to clustering unsupervised data mining techniques in the detection of procurement fraud?

ii. How is data mining used in fraud detection in procurement?

iii. Which data mining techniques can be used to unearth fraud in procurement

Significance of the study

This research study is significant because it will add value to the procurement fraud domain. This research will recommend the design that can be developed to curb fraud in the procurement sector using data mining techniques and technologies. The use of data analysis using data mining techniques, as unsupervised learning, for anomaly detection in procurement activities will serve as an automated tool for implementing procurement fraud detection processes that need in-depth verification. The results of this research will address the problem of fraud by proposing the automation of control tasks for fraud detection in the procurement process by comparing two unsupervised data mining techniques.

Scope

The research proposal covers the application of the two data mining techniques in all procurement process stages to help identify fraudulent activities.

Organization of the paper

The work is organized as follows:

1. Literature review of past works.
2. Analysis of data sets on the proposed solution
3. Results
4. Discussion of results

5. conclusion
6. recommendation.

CHAPTER TWO: LITERATURE REVIEW

Introduction

There is numerous information and resources on the study of fraud detection. However, Fraud detection is a very large field of study, whereby most of the papers and resources available consider detections as a primary tool. Nonetheless, Procurement fraud detection researches are limited; there aren't many studies in this particular field. This chapter focuses on fraud detection, outlier detection, and procurement fraud detection in the literature.

No matter the size of an organization, fraud is fatal to the very existence of any organization. Fraud has decapitated huge industries and corporations across all sectors. Over the years, fraud has been perceived as an internal issue. This norm is seen even in financial reports, where users are often made to believe that fraud is effectively deterred or hindered, and any attempt of fraud is detected and dealt with internally. Past works have presented anomalies in detection methods in procurement auctions using K-Means Clustering applied to a labeled refund transactions dataset. For instance, John et al., 2019 proposed a fusion approach for credit card fraud detection using a rule-based filter. As the previous work is appreciated, these works only deal with certain stages of procurement. The proposed work will combine anomaly detection in all stages of the procurement process. According to the PwC survey of 2020, procurement fraud was ranked the 6^{th} most prevalent form of corruption in the UAE, as shown in figure 1 below.

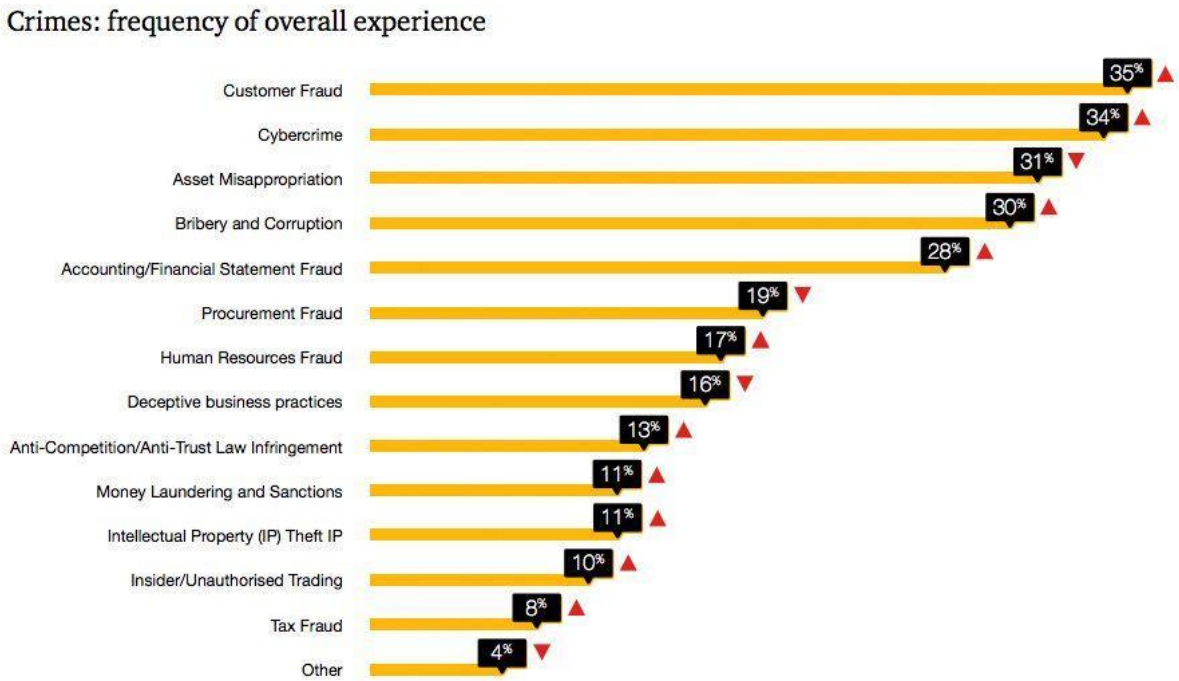

Figure 1: Fraud Prevalence in the UAE according to PwC 2020

Source:Consultacy-me.com

Numerous forms of corruption behavior are experienced in big companies or organizations. Preparation of complex financial reports, creating fictitious reports and creditors, following undue procedures in cost-cutting initiatives, and payment of expenses that have not been incurred are common forms of fraud easily disguised in financial reports (Hardinata,2021). Besides the auditing bodies, both internal and external, management bears the most responsibility in ensuring that fraud is detected and eradicated internally. It is also their responsibility to report such occurrences to the boards, shareholders, and the company's directors. Despite the need to have the top officials abiding by the set rules as an example, they are also deemed responsible for setting up fraud detection and solution mechanisms.

Fraud Detection Literature by Subject

Procurement Fraud and Detection

Procurement fraud has over many years been defined as obtaining something dishonestly, using pretense, avoiding an obligation, or causing a loss of personal or public property in a means that is directly or indirectly influenced by dishonest procurement. In a simple explanation, it is when the seller kickbacks a few percentages of the cost of the product to the buying company representative as a payment of a favored procurement decision he made. Compliance and fraud prevention is every Chief Financial Officers' greatest worry. In every company, procurement fraud prevention is one of the major priorities. Strong and systematic procurement will help in preventing procurement fraud and also attract external funding in the business. The procurement can happen in any part of the procurement process. However, auditing has to adapt to the growing amounts of data caused by digital transformation; to adapt to these vast growing demands, there is a need to implement machine learning and data mining, and analytic techniques into auditing (Nonnenmacher et al. 2021).

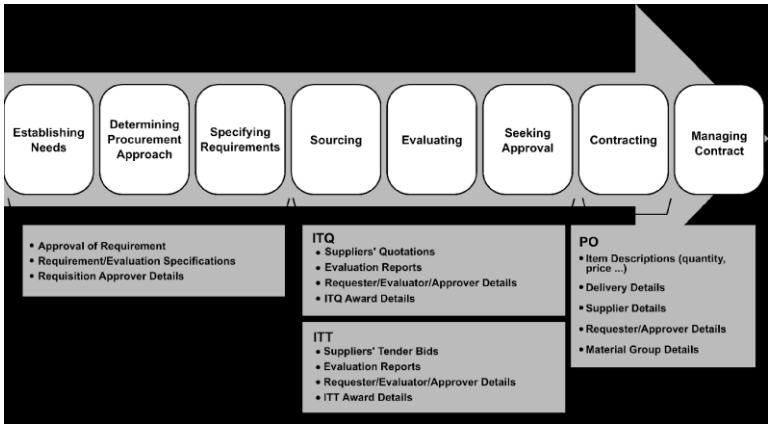

Figure 2 Procurement Workflow (Westerski et al. 2021).

From the analysis of various reasons for procurement fraud, the main reasons are classified into these categories;

- ***Financial needs:*** this is when the procurement person or company commits fraud because of the need to retain some percentage of the money to use the money to pay debt, gamble, or any other short-term needs and luxurious needs such as spending on vacations or buying expensive electronics.
- ***Perceive Opportunity:*** this is when the person identifies a hole in the company's procurement process and wants to use it to benefit themselves instead of covering the holes. In this situation, when the person committing fraud has a good understanding of the controls well or a total lack of controls on the procurement process, they have a wide opportunity for fraud.
- ***Rationalization:*** This is when the procurement officer tries to justify or explain their actions or thoughts. They do fraud because they are convinced that it is a one-time thing and it will never recur.

There are several frauds in the procurement, they include;

1) Employees colluding with suppliers (Velasco & Carpanese & Interian& Neto & Ribeiro, 2021).
2) The employees are setting up fraudulent companies to award themselves tenders.
3) Conflict of interest, occurring when someone controls the procurement process.

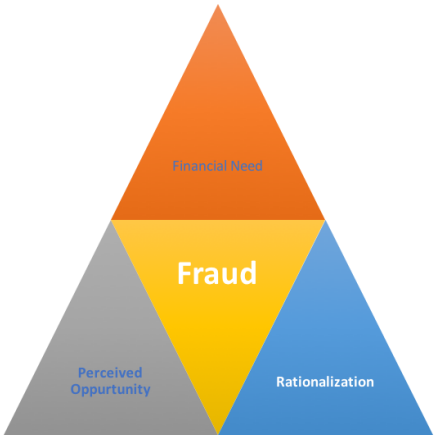

Figure 2 Procurement Triangle (Procuredesk.com)

Procurement Fraud Detection and Prevention

In the pre-contract award or pre tendering stage, fraud is hard to detect in this case. This is because, unless competitors regarding unusual contractor-client behavior fill complaints, it is likely to go unnoticed since some of the transactions and agreements can be contacted outside the company's premises (Velasco et al., 2021). The transaction or agreements also occur with or without the knowledge of the officers involved in the procurement processes. Some of these dealings include price-fixing between suppliers and the company officials to maximize the profit margins and even secure the tender or the business itself. This is particularly evident in the public entities winning the business, or the tender is partially specific on terms such as the lowest bidder and specialization (Velasco et al., 2021).

In the post-contract award stages, the fraud is more on the management part. This is because they handle most of the payment that is made on contracts. In this case, the loophole is in requisition, ordering, and delivery systems required to authorize the payments. Instances of submission of false invoices, overpayments, and similar fraudulent conduct are common at this procurement stage. The internal fraud organ successfully justifies the false invoices and overpayment to get the fraud proceeds from the suppliers (Velasco et al., 2021).

However, there are several ways to prevent fraud in procurement, they include;

i. Automation of control

The first automation system's design objective was to increase productivity; nonetheless, automation of control has helped fraud prevention. This is because all the automated purchasing systems have great transparency on procurement fraud prevention (Wersterski,2021). With the help of data mining, these purchasing systems can analyze a large amount of data, comparing them with other sources to identify trends in the procurement and prices valuations and inform the users on the best market prices. The procurers are discouraged from using the eyeball sampling technique in the compliance checks or when detecting fraud. This is because it will only cover a limited amount of orders compared to the total throughput of the organization (Westerski et al., 2021). The following is the summary of the capabilities of the automated purchasing systems;

a. The automatic purchase system helps in the creation of the purchase orders that ensure that all orders are approved at the right authorization procedures.
b. Purchase automation has ensured that every transaction has proof because of the production of order receipts and invoices.
c. The system also will ensure that the receipt and the invoices are matched. If not, it will raise the alarm.
d. The purchase system will also run the reports, provide historical purchasing information, and provide clear visibility on the expenditure.

ii. Being watchful

This is where the auditors, the owners, and the managers have a close eye on all the processes in the procurement department, by also doing market value research so that the procurement person

won't collide with the seller into hiking the prices of the resources to acquire, aiming for the kickback from the seller. This might include but is not limited to checking the employers' expenses and researching other possible sources of income. Their expenses should equal their annual income, and if not, there is fraud.

iii. Segregation of duty

This will ensure that one individual doesn't control all the procurement processes. A single person won't be able to make and approve purchases. Segregation of reduces has been proven to reduce procurement by a very large percentage. In most companies nowadays, two different departments handle the purchase orders and invoices. The procurement department will purchase orders and processes while the accounts department handles funds transfer and valuation. This reduces fraud

Credit Card Fraud and Detection

The crimes were committed using payment cards such as debit and credit cards. The aim is to make payments to obtain goods and services on transfer funds from the owner's account without their consent. Credit fraud can be authorized or unauthorized (Yogita et al., 2020). Authorized credit fraud occurs when the account owner transfers their funds to another account owned or controlled by a criminal. Unauthorized fraud is when the account owner doesn't approve the transfer of funds to another account. There are two types of credit fraud detection (Eswaran,2020);

i. On screening credit application

ii. Credit card transactions.

Several studies on case-based reasoning for eliminating credit approval propose a systematic data selection and support vendor machines web-based. Verma et al. 2020 proposed the use of distributed data mining.

Medical and Insurance Fraud and Detection.

Fraud in the medical and insurance sectors occurs when dishonest health care or insurance representatives or the insured/patient submit false or misleading information to benefit themselves or someone else. There are several techniques that scientists have implemented to try to detect and stop this fraud. MCPS can be used in health insurance. The Medical Cyber-Physical System allows incorporation and integration of medical implementations by use of Cyber-Physical System. They are used in hospitals and clinics as implantable and automated devices. (Omair, 2020). This system's cyber-attacks may encourage fraud in medical and insurance facilities.

Data Mining Techniques in Fraud Detection

Data is always an exemplary tool for solving observational, research, and experimental issues. Data analytics reflects on the structuring of the data to be usable and accessible to teams or individuals who require information about procurement in a company, for example (Diwakar et al. 2019). This essentially makes it easy to detect fraud and thus prevents it from happening. On the other hand, data mining is the process by which raw data is turned into information. In this method, the software is used to synthesize patterns of the procurement processes in an organization so that loopholes can be identified and the sourcing of general information (Diwakar et al., 2019).

Data mining involves evaluating and analyzing huge sets of data to establish meaningful trends and patterns from which information can be extracted (Sanches,2021). The process involves data collection. The next step involves the storage and management of data in servers or the cloud. Depending on the intended use, the data is organized, and software is applied to generate actionable data. The following are analyses of data mining techniques used in fraud detection;

Tracking patterns

This is a simple data mining technique whereby the user goes through the data sets to identify patterns in the data. It also checks on the inadequacies that may arise and the frequency and intervals at which certain phenomena occur. This technique is not only good in establishing product-market trends, but it can be a useful tool in finding inconsistencies and hence fraud in procurement (Diwakar et al., 2019).

Classification

This is more complex and detailed compared to tracking patterns. In this technique, data with a certain attribute are put together in categories that are discernable to the person interested in deducing information from the data (Verma,2021). Association, on the other hand, is more related to tracking patterns. It, however, goes a notch higher to find a correlation between events and or attributes (Diwakar et al. 2019).

Outlier detection

Outlier detection seeks to aim at recognizing patterns that are over-changing in a data set. This is where absurd trends and anomalies are detected and pointed out for further action. Clustering, on the other hand, is similar to classification. Clustering groups data to numerous sets based on their similarities. Other data mining techniques include regression and prediction, and both are used depending on the circumstances (Diwakar et al., 2019).

Clustering

Clustering forms grouped data points into clusters so that similar objects or characters are easily identifiable in subsets. The clusters are the basis a company or business uses to make informed decisions about their products and the target consumer. It isn't easy to solve as an unsupervised machine learning tool in terms of the relevant algorithm (Maia et al., 2020). From

the basics, the parameters that define the strength of the data and the proposed outcome are intra-cluster similarity.

Reasons and applications of clustering in data mining

Cluster analysis has an extensive footprint in data mining as it canvases through different fields and has a variety of applications. The range begins from imagery, mobile communications, economics, medicine, and computational biology. It, however, has the limitation that it cannot be standardized. It may give the best results with a single type of data set but may not be the best when there are varying data sets, and hence it is not the best in all real circumstances (Maia et al., 2020).

The scalability of the clustering requires a similar or approximate scale to the complexity order of an algorithm as per the boost of the number of objects in the algorithm. For example, when 20 folds raise the number of objects in a cluster, the proportional time should increase approximately 20 times. Clustered data should also be understandable and therefore usable in the proposed projections or predictions. Even though it is seen as the weakness of clustering analysis, algorithms should interpret and generate different data attributes and, at times, deal with noisy data (Maia et al., 2020).

Data tracking in applications and other systems will be key in fraud detection and tracking for this research. Since the research will focus on the products from the target companies and the people, for example, detecting certain behavior and traits will be clustered or associated with the fraud. Hence, their detection will prompt suspicion of the concerned individuals.

Association Components

Association uses rules and statements such as "if-then," which majorly display the probability of relationships between data items. It is a data mining technique mainly used in the correlation of financial and medical data sets. Medical practitioners use it to assist in the diagnosis of patients (Hussain et al., 2018). Using known symptoms and those portrayed by a particular patient, it is easy for doctors to determine the illness using conditional probability. In retail and procurement setups, the association is used to group the customer demand and purchasing patterns and hence help the business adjust their sales and marketing strategies (Hussain et al., 2018).

In data mining and association, in particular, some rules guide the development of patterns at the basic level that machines can interpret and analyze the data in the long run. The antecedent mainly guides association (if) and consequent (then) rules. The first is normally an item found within the data, while the latter is an item found in combination with the antecedent. A criterion is then created in the search interface based on the frequency of the two rules, if-then, which give confidence and support to a purported behavior or relationship (Hussain et al., 2018).

Association rules, algorithms, and the use in fraud detection

There are popular algorithmic rules and variations such as Apriori, SETM, and AIS. Candidate item set is generated using a large item set in the Apriori of the previous pass. All item sets with sizes larger than one of the previous passes are joined. The others are large and are then deleted (Wang and Zheng, 2020). The STEM algorithm scans databases by scanning candidates item sets and, in the end, account themselves as a candidate in a sequential data structure. For this project, AIS algorithms count item sets and generate the scans (John,2019). It is relevant in this case as it determines large item sets contained in a transaction.

The goal is, therefore, to analyze and predict customer behavior. They are also important in analyzing market baskets and, in our case, establish the trends in a particular company that help in fraud detection. Companies' trends may be answered through the association rules of “if "and "then." Their occurrences and predictions also form the basis of the parameters of support and confidence depending on the situation (Wang and Zheng, 2020).

Supervised Data Mining Methods

The supervision has also been implemented in fraud detection and it involves the techniques of grouping objectives and projection objectives. It involves the use of the traditional statistical methods. Examples of traditional statistical methods include;

i. Neural networks
ii. Bayesian network
iii. Support vector Machine
iv. Discriminant analysis
v. Regression analysis

Accuracy in grouping of records and certainty in the truth are needed by these supervised data mining approaches.

The supervised data mining methodologies include the Support Vectors Machine, neural networks, and genetic algorithms. These methodologies have been applied to expose fraud in the procurement process in the procurement sector and other sectors, including healthcare and transport.

Unsupervised Data Mining Methods

The working methodology of the unsupervised data mining methods involves evaluation of an individual's claim aspect and its characteristics in respect to the other claims, after the evaluation

process, the association between these claims and the way they are associated is figured out. These methods are used for analysis and characterization which include the segmentation techniques. The segmentation techniques used in the unsupervised data mining methods include anomaly detection, clustering and association rule. With the use of unsupervised data mining techniques, it is possible to identify and remove patterns and association rules amongst the records, identify any form of irregularities or less or more equal records. Among the many unsupervised techniques, the outlier detection, association rule, and clustering are a few techniques that have been used to detect and expose fraud in the procurement.

In his articles dating the year 2020, Rahul gives a deep understanding of the working methodology of the unsupervised data mining methods; it involves evaluating an individual's claim aspect and its characteristics in respect to the other claims. After the evaluation process, the association between these claims and their associates is figured out. These methods are used for analysis and characterization, which include segmentation techniques. The segmentation techniques used in the unsupervised data mining methods include anomaly detection, clustering, and association rule. With unsupervised data mining techniques, it is possible to identify and remove patterns and association rules amongst the records and identify any irregularities or less or more matching records. Among the many unsupervised techniques, outlier detection, association rule, and clustering are a few techniques that have been used to detect and expose fraud in procurement (Vardhani,2019).

Rahul (2020) presents two classifications of data mining models in data mining; these are the linear and non-linear models. The linear entails mathematical modeling and approaches or patterns that consider the linear association parameters amongst the constants or several parameters under study. The non-linear model does not take on or accept the linear association

between the parameters and constants under study. The linear models are praised as stable and straightforward models; however, they pose evident possible difficulty.

According to Rahul (2020), This classifier, the Naïve Bayesian, implements the Bayesian theorem in naïve objectivity. The Naïve Bayes algorithm uses the Bayes Theorem to perform its classification. The Naïve Bayes is based on applying the Bayes Theorem by choosing the independent assumptions amongst the features. The Bayes theorem is used for classification by calculating the probabilities of the parameter values of the classification groups individually and using these probabilities to predict the undefined instances. If the data dimensions entered are higher, the Naïve Bayes classifiers method is the best suited for its analysis. According to the Naïve Bayes classifiers, the attributes and elements are hypothetically independent, and there is no interdependence between the attributes. The pros of the Naïve Bayes are its ease and simplicity of implementation; it doesn't demand much data training. Again, Naïve Bayes can handle both discrete and continuous data. However, suppose the data in the test vary from the category of the test data set. In that case, the NAÏVE Bayes will assign its probability a zero won't be able to make any predictions regarding the situation in the new data set (Haofan, 2016).

In addition, there is another unsupervised data technique discussed by Rahul (2020). This is the Decision Tree. The research he did with his colleagues presents the Decision tree as a support tool that comprises the roots, the stem (non-leaf) node, and the leaf node, and it has a tree-like structure. The non-leaf node represents the analysis of the attributes, and the branch and leaf node represents the outcome of the analysis and the class babel, respectively. The methodology of work adopted in the decision tree is by splitting the data into subcategories repetitively. Each subcategory will comprise data that is identical in states of the intended variables. The dividing formula is selected in a criterion chosen that properly divides the dataset given. The decision comprises the three

algorithms. These algorithms used the information gain, the Gini Index, and the Gain Ratio as their standard sampling measure (Anamika et al., 2018).

Data Mining Classification Techniques

There are two classifications of data mining models in data mining, these are the linear and non-linear models. The linear entails mathematical modelling. approaches or patterns that take into account the linear association parameters amongst the constants or several parameters under study. The non-linear model does not take on or accept the linear association between the parameters and constants under study. The linear models are praised as stable and straightforward models, however, they pose clear possible difficulty.

Naïve Bayes

This classifier, the Naïve Bayesian, is centered to implement the Bayesian theorem, in naïve objectivity. The Naïve Bayes algorithm uses the Bayes Theorem to perform its classification. The Naïve Bayes is based on applying the Bayes Theorem, by choosing the independent assumptions amongst the features (Peling,2017). The Bayes theorem is used for classification by calculating the probabilities of the parameter values of the classification groups, individually, and using these probabilities to predict the undefined instances. If the dimensions of the data entered is higher, the Naïve Bayes classifiers method is the best suited for its analysis (Kiran,2018). According to the Naïve Bayes classifiers, the attributes and elements are hypothetically independent and there is no interdependence between the attributes. The pros of the Naïve Bayes are its ease and simplicity of implementation; it doesn't demand much training of the data. Again, Naïve Bayes can handle both discrete and continuous data. However, if the data in the test vary from the category of the test data set, the NAÏVE Bayes will assign its probability a zero admit won't be able to make any predictions in regards to the situation in the new data set (Husejinovic,2020).

Decision Tree

This classifier is a support tool that comprises the roots, the stem (non-leaf) node and the leaf node, it has a tree-like structure. The non-leaf node is the representation of the analysis of the attributes and the branch and leaf node represents the outcome of the analysis and the class babel respectively(Campus,2018). The methodology of work adopted in the decision tree is by splitting the data into subcategories, repetitively, in a way that each subcategory will comprise of data that identical in states of the intended variables. The dividing formula is selected in a criteria chosen that properly divides the dataset given. The decision comprises of the following algorithms;

i. CART,

ii. ID3

iii. C4.5

These algorithms used the information gain, the Gini Index and Gain Ratio as their standard sampling measure (Khare,2020).

Research Studies Data Mining in Fraud Detection

There are quite a number of research studies done relating to the use of data mining techniques and technologies in fraud detection. The following include the studies that reviewed and studies, (Richard and Taghil,2018). These devoted researchers are acknowledged herein for their great contributions by investigating the fraud discovery techniques using data mining and machine learning.

Richard & Taghil in their research in 2018, proposed the implementation of machine learning techniques by using the available fraud claim as data set and labels for the identified fraudulent activities among the medical practitioners. With this proposition, they went ahead to demonstrate the usefulness of using machine learning with under-sampling to detect fraud in the health sectors.

Maia (2020) proposed the implementation of an artificial neural network which trains with insurance dataset which will anticipate fraud and fraudulent actions in the insurance sector. Maia (2020) went ahead and created this artificial neural network and it contributed to fraud protections in the institutions.

Summary

In this chapter, I analyzed the key definition of terms and topics associated with research. The definition of procurement has been reviewed in depth. The types of fraud have also been analyzed, and examples are given, i.e., procurement fraud, credit card fraud, and medical insurance fraud. I have conceded and highlighted the role of data mining and machine learning in procurement and fraud detection. This has been with the analysis of various data mining techniques that are used in Fraud procurement. These techniques include; tracking patterns, classification, outlier detection, clustering, and association components techniques. In summary, this chapter has facilitated the research in acquiring the necessary knowledge to aid this research during data acquisition and analysis.

Unsupervised Data Mining Methods

In his articles dating the year 2020, Rahul gives a deep understanding of the working methodology of the unsupervised data mining methods; it involves evaluating an individual's claim aspect and its characteristics in respect to the other claims. After the evaluation process, the association between these claims and their associates is figured out. These methods are used for analysis and characterization, which include segmentation techniques. The segmentation techniques used in the unsupervised data mining methods include anomaly detection, clustering, and association rule. With unsupervised data mining techniques, it is possible to identify and remove patterns and association rules amongst the records and identify

any irregularities or less or more matching records. Among the many unsupervised techniques, outlier detection, association rule, and clustering are a few techniques that have been used to detect and expose fraud in procurement (Vardhani,2019).

Rahul (2020) presents two classifications of data mining models in data mining; these are the linear and non-linear models. The linear entails mathematical modeling and approaches or patterns that consider the linear association parameters amongst the constants or several parameters under study. The non-linear model does not take on or accept the linear association between the parameters and constants under study. The linear models are praised as stable and straightforward models; however, they pose evident possible difficulty.

According to Rahul (2020), This classifier, the Naïve Bayesian, implements the Bayesian theorem in naïve objectivity. The Naïve Bayes algorithm uses the Bayes Theorem to perform its classification. The Naïve Bayes is based on applying the Bayes Theorem by choosing the independent assumptions amongst the features. The Bayes theorem is used for classification by calculating the probabilities of the parameter values of the classification groups individually and using these probabilities to predict the undefined instances. If the data dimensions entered are higher, the Naïve Bayes classifiers method is the best suited for its analysis. According to the Naïve Bayes classifiers, the attributes and elements are hypothetically independent, and there is no interdependence between the attributes. The pros of the Naïve Bayes are its ease and simplicity of implementation; it doesn't demand much data training. Again, Naïve Bayes can handle both discrete and continuous data. However, suppose the data in the test vary from the category of the test data set. In that case, the NAÏVE Bayes will assign its probability a zero won't be able to make any predictions regarding the situation in the new data set (Haofan, 2016).

In addition, there is another unsupervised data technique discussed by Rahul (2020). This is the Decision Tree. The research he did with his colleagues presents the Decision tree as a support tool that comprises the roots, the stem (non-leaf) node, and the leaf node, and it has a tree-like structure. The non-leaf node represents the analysis of the attributes, and the branch and leaf node represents the outcome of the analysis and the class babel, respectively. The methodology of work adopted in the decision tree is by splitting the data into subcategories repetitively. Each subcategory will comprise data that is identical in states of the intended variables. The dividing formula is selected in a criterion chosen that properly divides the dataset given. The decision comprises the three algorithms. These algorithms used the information gain, the Gini Index, and the Gain Ratio as their standard sampling measure (Anamika et al., 2018).

CHAPTER THREE: METHODOLOGY

This chapter covers the research philosophy, design, and data collection and analysis that covered.

Introduction

The project's goal was to train an unsupervised learning model that separates anomalies from the data for their consequent human analysis to determine if fraud occurs. The techniques of data mining and processing will have applied as illustrated in figure 3 below.

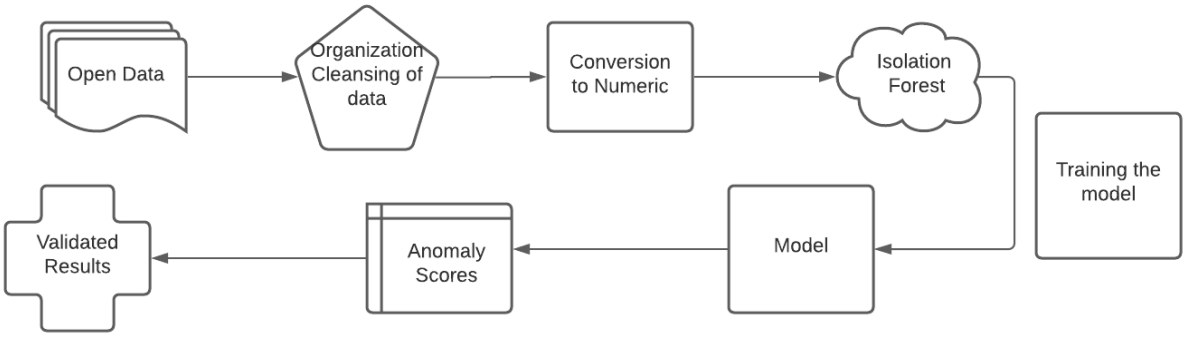

Figure 3: Proposed data mining workflow solution

Dataset

The dataset used in this research was obtained from kaggle.com about recent transactions that have occurred in an unknown firm, due to discretion, of the United Arab Emirates, and as such is a secondary source rather than primary. It contains transactions made by the firm, some of which are fraudulent as required by the project to prove the functionality of the algorithm. The fraud dataset is pre-processed to ensure that it is compatible with the methodologies and algorithms that will be employed (Haofan, 2016). Clustering is a method for categorizing data. In this case the data must be very comparable inside a cluster but distinct across clusters. In this research, the sourced data from losses due to theft, fraud and vandalism in the United Arab Emirates available in the world bank dataset (data.worldbank.org).

The data is made up of records. These things are called attributes in a record like this one. In theory, a record can't be faked. On the other side, a collection of records representing a transaction might be fraudulent. Take, for example, a record that shows how much money was paid to supplier X. Another record says that X paid less money for a previous purchase order than on the invoice. Separately, both of these records aren't fake. The only way to judge this transaction as illegal is when it's all done at the same time, records are used to keep track of what people do. The frauds are the main thing that want to find out about. The only way to find out if someone is being dishonest is to look at how they act. Observing how employees act can be done by looking at things that go with their behavior. People have certain traits that are based on the traits of transactions, which are based on the traits of records. To obtain even more insightful data, additional qualities are added to those currently provided. To get a suspicion score, look at the attributes that are at the top of the list. This score gives an idea of how likely it is that an employee will be dishonest in the future. This engineering assumes that a fraudster's conduct is very different from that of a trustworthy worker's. If not, there will be no distinction between qualities indicating a fraudster's conduct and those identifying a regular employee. Thus, no suspicion scores will be very different from the other scores. Given a collection of n data points, the purpose is to cluster them into k clusters, with the clusters positioned precisely so that the distance between the data points and the cluster is as little as possible. The data point distance from the cluster centroid is determined as follows (Anamika et al., 2018):

$$d\,(cluster\ centroid) = \sqrt{(X - X_c)^{\,2} + (Y - Y_c)^{\,2}}$$

Algorithm:

Step 1: K cluster centroids (number of clusters) initialized.

Step 2: Set up a cluster with the smallest distance between each point's initialized centroids.

Step 3: Calculate the newly created clusters' centroids.

Step 4: Perform procedures 2 and 3 till the centroids remain stationary, i.e. they converge.

K-modes Clustering

The concept of K-modes clustering has been frequently used in scenarios involving categorical data. For categorical data, statistical inference measures such as mean cannot be computed, and k modes overcome this constraint of the k-means algorithm precisely. For categorical data, a dissimilarity metric such as hamming distance may be utilized (Haofan, 2016).

$$d(x, y) = \sum_{i=1}^{n} (Z_i, Q_i)$$

Algorithm (Singh, R., & S, M., 2018).

Step 1: Initialisation of k clusters by selecting k starting modes which may or may not exist in the analyzed dataset.

Step 2: Using the dissimilarity measure, determine the object's resemblance to the cluster modes that are accessible. If object!=cluster mode, dissimilarity=1. If (object==cluster mode), dissimilarity=1-nrj/total.

Step 3: Assign the evaluated data point to the cluster with the lowest dissimilarity score.

Step 4: Update the cluster modes to reflect the newly added cluster objects.

Clustering on a Hierarchical Scale

Hierarchical Clustering divides data points into clusters in a hierarchy. Clustering may be accomplished using either the 'agglomerative technique of hierarchical clustering' or the 'divisive method of hierarchical clustering (Anamika et al., 2018). The Agglomerative technique of hierarchical clustering, often known as the bottom up approach, entails assigning each data point or data item to its own cluster. After then, depending on the items' similarity, their separate clusters are combined to create a larger cluster. This procedure is done until all related items are grouped together (S et al., 2018). The divisive, or top-down, form of hierarchical clustering entails grouping all data points or data objects into a single cluster. The data points are then clustered according to their dissimilarity. This procedure is continued until all dissimilar data points are assigned to distinct clusters. Given a collection of k clusters and a KXK matrix with the distances between them (Haofan, 2016).

Algorithm (Singh, R., & S, .. M., 2018)

Step 1: Assign each data point to its own cluster. The distance between clusters is the same as the distance between data points.

Step 2: Identify a pair of clusters with the shortest distance between them and merge them into a single cluster.

Step 3: Find the separation between the two clusters.

Step 4: Repeat the steps 2 and 3 until all data points are clustered together.

Various hierarchical clustering approaches include the following:

1. ***Single linkage:*** The smallest distance between two clusters.
2. ***Complete Linkage:*** Denotes the greatest distance between data points in one cluster and data points in another cluster.
3. ***Centroid:*** Denotes the distance between the data points corresponding towards the centroids of clusters considered.
4. ***Ward:*** This variable is the sum of the squares of the two clusters' data points.

Medoids-Related Partitioning

Outliers are detected via the k-means algorithm. A more effective approach is offered by the clustering technique known as partitioning around medoids (PAM). The clusters are represented here as medoids rather than centroids. It is capable of handling any data types, including continuous (Haofan, 2016).

AlgorithmSingh, R., & S, .. M., 2018).

Step 1: Pick k medoids at random to symbolize k clusters.

Step 2: Measure the data points' distance from either the cluster medoids.

Step 3: Assign the piece of data to the cluster closer to the medoid in question.

Step 4: Calculate the total distance between the data points and the medoids by adding the distances between the data points and the medoids.

Step 5: Replace the current medoid with a position that is not a medoid.

Step 6: Reassign each data point to the cluster of the nearest medoid.

Step 7: Add up all of your expenses and determine your final budget.

Step 8: If the computed total is less than the previous medoid, then the new point is retained as the next medoid.

Step 9: Continue Steps 5–8 until the medoids have converged.

Self-Organizing Maps (SOM).

The most often used neural network methodology, Self-Organizing Maps (SOM) is a competitive learning strategy. An unsupervised learning method, SOM does not rely on human supervision and has no idea about the qualities of the input data it uses to learn. The self-organizing mappings do not require the input data to be labeled. It performs a mapping of higher dimensions to map units (while preserving the topology). The spots that are closer together are grouped together as nearby map units. The neighboring map units create a lattice, allowing high-dimensional mapping to a plane. The SOM keeps track of the distance between two points in question. SOMs are capable of generalization and hence may be utilized for cluster analysis. Self-organizing maps generate extracted features from continuous space and convert it to discrete space (Haofan, 2016).

Algorithm:

Step 1: Pick random weights for the map's initialization.

Step 2: Sample the input space by selecting a subset of vectors.

Step 3: Select the neuron whose weight is most closely aligned with the given vector.

Step 4: Update the equation.

Step 5: Repeat steps 2–4 until the SOM is constant or remains constant.

Research Design

Quantitative experimental research was chosen as the best approach for this research, gathering, assessing, valuing and analyzing data.

Business Understanding

This research used primary and secondary sources to understand the business operation, setting, and procurement processes. The research aimed to understand the goals and necessities from the procurement perspective regarding how rogue employees are undertaking fraud. This business understanding methodology helped in understanding the data mining techniques that can be implemented in this sector (Haofan, 2016).

Research sampling

I have sampled the number and types of companies to take part in the research. The engaged population are questioned based on the topic and the research objectives. The survey was conducted with the managers and employees in their procurement sector in identified companies in Dubai, UAE. The size of the sample is 100. In this survey, a questionnaire is provided to the sample participants, filling the paper and returning the filled questionnaire to the researcher for analysis. By doing effective sampling and survey, this research was well-structured and effective.

Code

Importing Libraries and Dataset

It begin by importing the basic python libraries that facilitate fraud detection capabilities and a sample dataset to work on as shown below:

```
[6]:  import pandas as pd
      import numpy as np
      import matplotlib.pyplot as plt
      import seaborn as sns

[7]:  pd.options.display.float_format = '{:.2f}'.format

[8]:  d1 = pd.read_excel('card transactions.xlsx')

[9]:  d1.shape
```

Information Display

It display information of the imported dataset to verify it's one that can work with to yield the desired results as shown:

```
d1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 96753 entries, 0 to 96752
Data columns (total 10 columns):
Recnum               96753 non-null int64
Cardnum              96753 non-null int64
Date                 96753 non-null datetime64[ns]
Merchnum             93378 non-null object
Merch description    96753 non-null object
Merch state          95558 non-null object
Merch zip            92097 non-null float64
Transtype            96753 non-null object
Amount               96753 non-null float64
Fraud                96753 non-null int64
dtypes: datetime64[ns](1), float64(2), int64(3), object(4)
memory usage: 7.4+ MB
```

Further describe the dataset statistically based on previous records of frauds summatively.

```
d1.describe()
```

	Recnum	Cardnum	Merch zip	Amount	Fraud
count	96753.00	96753.00	92097.00	96753.00	96753.00
mean	48377.00	5142201786.82	44706.60	427.89	0.01
std	27930.33	55670.84	28369.54	10006.14	0.10
min	1.00	5142110002.00	1.00	0.01	0.00
25%	24189.00	5142152067.00	20855.00	33.48	0.00
50%	48377.00	5142195612.00	38118.00	137.98	0.00
75%	72565.00	5142246089.00	63103.00	428.20	0.00
max	96753.00	5142847398.00	99999.00	3102045.53	1.00

Ethical consideration

The ethical considerations were not neglected in the whole process of this research as the main principles that covered the process were confidentiality, fairness, honesty and unbiasedness. The researcher upheld the openness principle throughout this process. Confidentiality of the participants was not interfered with; their names and identities were and will not be revealed to anyone, and any sensitive information received were destroyed after compiling the data collected. Fairness and rationality were among the fundamental principles during this research, and it was always maintained until the end of the survey. The researcher did not, by any means, force the participants in the survey to submit the answers; they all did out of the will (Haofan, 2016).

Methodology limitations and challenges

Data collection and the intended training of the unsupervised model require time and a one on one interaction with the employees of these companies and the general public because the objective of an unsupervised learning algorithm is to derive insights from massive amounts of new data (Zangirolami,2018). It is also important to mention that the UAE and the world are still struggling

to get out of the grip of the Covid-19 pandemic. If the social distancing rules are to be adhered to, the study will require minimum interactions and interviews that may be conducted. When all factors are considered, it is better to do the training, the interviews and the questionnaires through online platforms. This essentially means that there will be a limitation of biases in such circumstances whereby the people may not be genuine as compared to a one on one or physical interaction.

CHAPTER FOUR: PRIMARY AND SECONDARY DATA

Introduction

In this research paper, all the technical and non-technical aspects have been mentioned. It was necessary to collect primary and secondary data. This data would then be analyzed and found recorded, basing the results on the objective of this study (D'allerto,2021).

To collect the required data, I surveyed the procurement companies and organizations with elaborate procurement departments in all the identified sectors in the UAE. The primary data was collected using the survey questionnaires and interviews. Apart from using the primary data collection techniques, the secondary method of data collection was employed in this data collection process. The data was fetched from different books, journals, websites, and articles relevant to this study and all of them were acknowledged herein.

The thematic analysis and quantitative analysis of data are also covered in this chapter. The thematic of secondary data is conducted, and secondary data is evaluated. The survey answers were classified into different groups, and these were the approved, denied the cancelled and ambiguous. The data filtering followed the data preparation process. In this process, the data with only complete and helpful information was selected from the collected many answers. Repeating claims were removed, the claims with zero amount and missing columns were also removed. After selection, the information presented below were our findings..

Secondary Sources

As mentioned in the previous chapters, several accredited and reliable secondary sources were used in this research. Several books, journals and articles were used to source the necessary information for this research, various reports and resources from procurement groups around the UAE.

CHAPTER FIVE: DATA ANALYSIS

Implementations

R 3.3.2 is used to carry out the planned task. The data set utilized is titled sales fraud detection,' and it has five attributes (Haofan, 2016): ID, Prod, Val, Quant, and Insp. The following are the characteristics:

a) ID: Identifies the salesperson. b) Prod: Identifies the product.

c) Val: This field contains the transaction's reported value.

d) Quant: This field contains the amount of a certain product.

e) Insp: This field contains a report of three potential values, namely acceptable, fraudulent, and unknown.

The label attribute, insp, from the dataset in this case was categorical. As a result of this pre-processing, it may apply the data to a variety of classifiers with greater efficiency. Unkn is encoded as a '1' to indicate that the state of the transaction is not known, while Ok is encoded as a '0'. Fraud is encoded as '2', indicating that the transaction involves a detection of fraud (S et al., 2018)

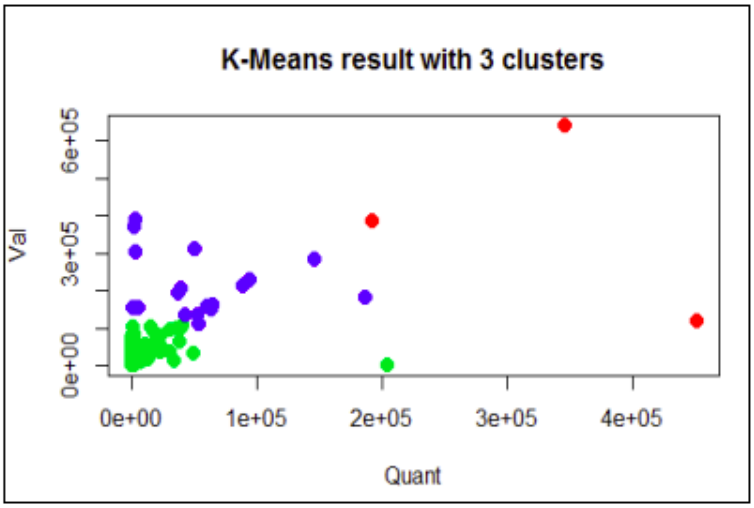

Figure 4: K-Means Result

Techniques for Clustering.

Clustering via K-means:

In Figure 4, begin by assuming that the parameter k equals three. The following is the outcome: The green dots indicate non-fraudulent transactions, whereas the blue dots indicate transactions for which the status of fraud or not is uncertain. The red dots indicate instances of fraud. The above procedure achieves a precision of 70%. After that, I use the elbow approach to determine the ideal cluster size for the current circumstance. After eight clusters, the approach demonstrates that the graph remains constant. Thus, as seen in Figure 5, I employ eight clusters to represent our dataset (Haofan, 2016).

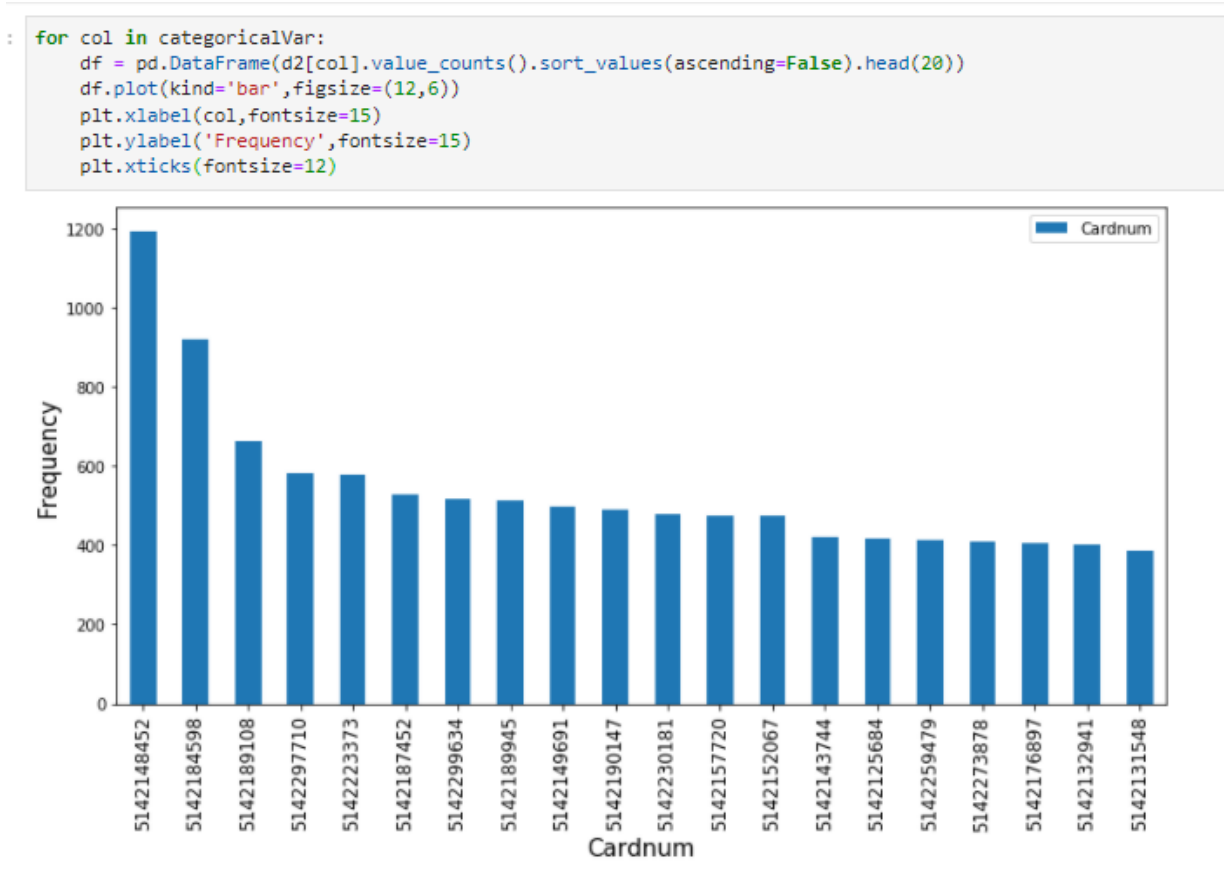

Figure 5: K-Means Performance

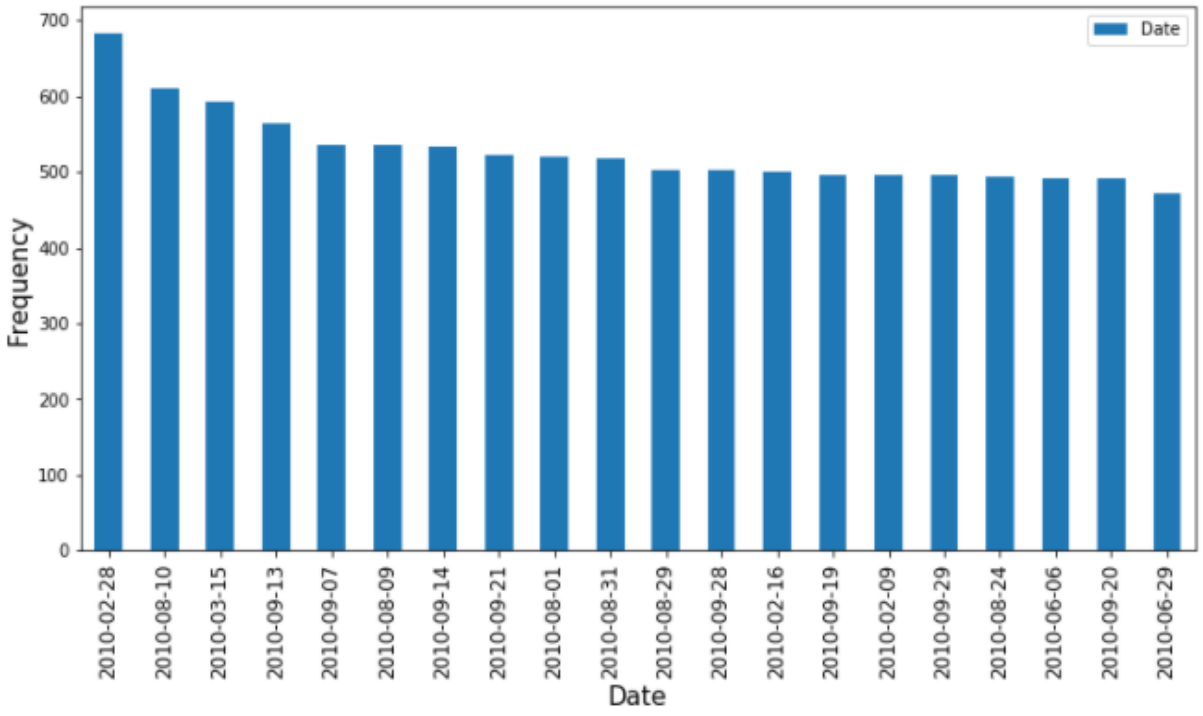

Figure 6: K-Means Performance

Clustering Via K-Mode:

As in Figure 6, the actual transaction is represented by blue, grey, black, and cyan. This approach has a 94.5 percent accuracy rating. Using the k-modes clustering approach, I can see that the black dots in the preceding figure belong to the group of non-fraudulent transactions. As depicted in Figure 7, the red dots indicate transactions that are unsure about their status, while the green dots reflect the cluster of transactions that are fraudulent (Anamika et al., 2018).

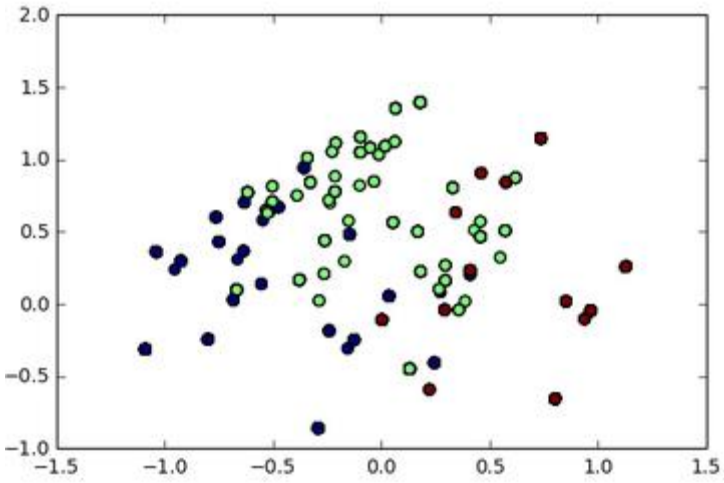

Figure 7: Hierarchical Clustering Plots

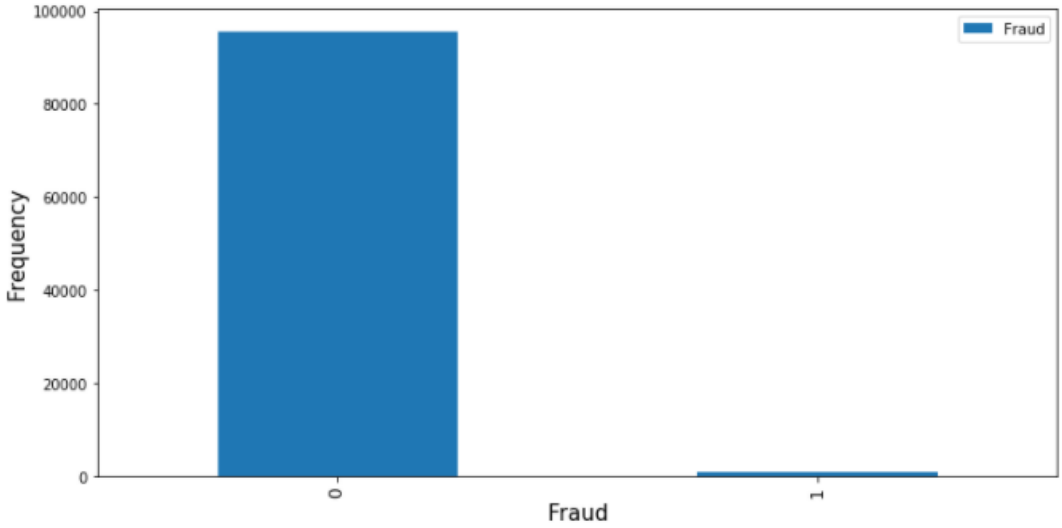

Figure 8: Frequency vs Fraud Visualization

By using hierarchical clustering, as it can see that the dark blue dots reflect transactions that are valid, i.e., neither fraudulent or unlikely to be fraudulent. In Figure 8, the light blue dots indicate transactions that are extremely likely to be fraudulent, while the moderate blue represents transactions that are likely to be fraudulent.

For example, a plus sign indicates “fraud” values, a triangle indicates “unkn” values, and a circle indicates “ok” values, which reflect non-fraudulent transactions.

Clustering via Medoids Partitioning

K-medoids clusters are improved using the PAM technique, which analyzes the full dataset to look for the most promising candidates for clustering. The PAM technique is called Partitioning Around Medoids (PAM). As a result, PAM outperforms other algorithms when it comes to medoids. To speed up the swap phase of PAM in k-medoids clustering, this study offers a parallelization of PAM on the GPU. Using shared memory, a reduction method, and thread block configuration optimization, the parallelization scheme maximizes occupancy.

```
plt.figure()
plt.plot(list(sse.keys()), list(sse.values()), 'bx-')
plt.xlabel("Number of cluster")
plt.ylabel("SSE")
plt.show()
```

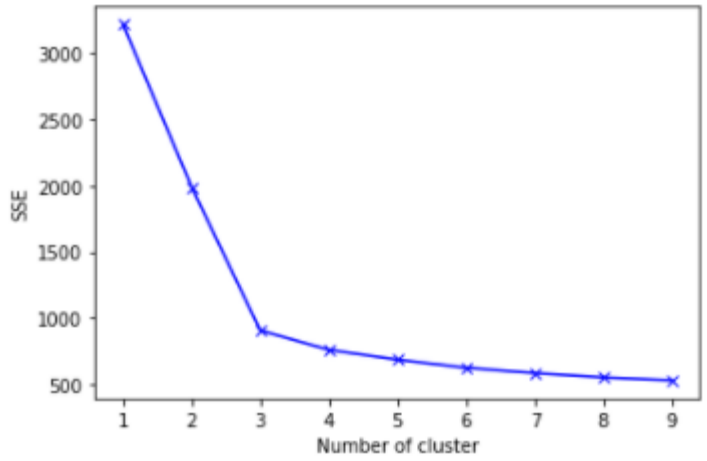

Figure 9: Medoids Partitioning

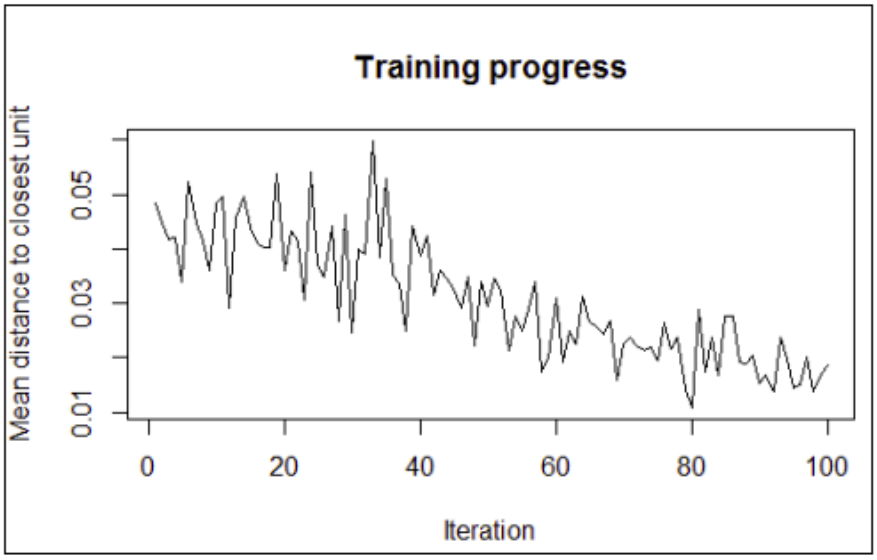

Figure 10: Self Organising Maps (SOM)

Clustering via SOM:

As seen in Figure 10, the mean distance between comparable data points decreases as the number of repetitions increases. In the picture above, the neighbor distance plot indicates that the points in red are clustered together and hence have a smaller distance between them, while the dots in yellow reflect data points that are farther apart (V et al., 2017).

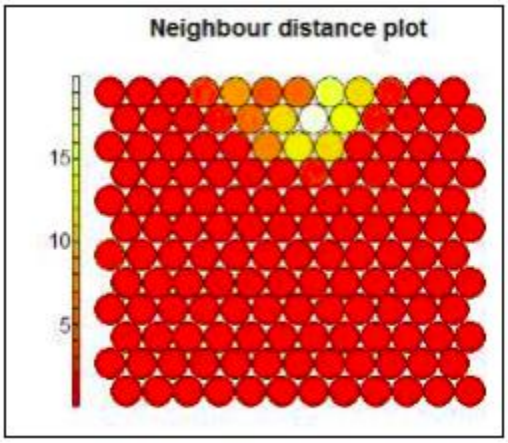

Figure 11: Graphical Representation 1 of Transactions

The red dots on the counts graphs represent transactions that are either fraudulent or not. The orange ones indicate legitimate transactions. Grey units are used to represent empty units. After using the SOM approach, clustering is conducted. Simultaneous transactions are combined. The self-organizing map technique has a 99.2% accuracy rate. Thus, the following summarizes our findings about the strategies used (V et al., 2017):

Table 4.1: Clustering Techniques Comparison

Clustering Techniques	**Percentage Accuracy**
Partitions Around Medoids	94.75
SOM	99.35
Hierarchical Clustering	98.80
K-Modes	94.80
K-Means	94.50

As seen in the table above, self-organizing maps have the maximum degree of accuracy at 99.2 percent. Additionally, hierarchy produces excellent outcomes at 98.8%. The k-means method achieves a 94.5 percent accuracy, whereas the k-modes approach achieves a 94.8 percent accuracy by dividing around medoids (Anamika et al., 2018).

Similarity in DISC.

It is shown in this research that the efficacy of a ranking algorithm depends on the similarity measure. Similarity is computed using distributions of frequently occurring feature values. Similarity measures of this sort may correct for unusual nominal values, which adds more information to the representation of the link between data instances. This may enable the detection of novel patterns in comparison to simpler feature-based similarity techniques.

SRA Performance Evaluation.

ROC curves provide a class skew independent performance metric for evaluating the proposed SRA's ability to identify anomalies. The positive class corresponds to the anomalous cases, whereas the negative class corresponds to the normal cases. Consider the total number of positive and negative examples as n+ and n correspondingly. False positive rate is shown versus real positive rate when the threshold level changes. The genuine positive rate (or sensitivity) is one assessment criteria, whereas the false positive rate (or specificity minus one) is another (V et al., 2017).

False-positive and true-positive boundaries are represented by the ROC curves, which show how varying threshold values affect the trade-off between benefits and costs. Thus, combining the two goals does not need previous information. A ROC curve that outperforms another gives a superior solution at any cost point, as shown by a larger area under the curve (AUC). Probability of ranking positive class samples above negative class samples is calculated by using the AUC (Average Utility Classifier) metric, which takes into account random sampling from both classes. Thus, the ROC curves and AUC may be used to assess an algorithm's performance on specific data sets. Following that, will utilize AUC and ROC curves to evaluate the performance of various techniques (S et al., 2018).

Additional Validation of the SRA Ranking.

To further validate the SRA rating obtained, I explore the observed cluster structure. I investigate crucial traits that distinguish highly rated claims from the rest. If the difference from the majority is due to characteristics that are prone to deception, this lends credence to the derived rating. Consider clusters discovered using the Hamming kernel with a value of 0.8.

Like Subplot (a) in Figure 7, but with clear labels for first and second non-primary eigenvector clusters. As seen in the plot and table, 92 percent of fraudulent instances really fall into clusters 4, 6, and 7. Additionally, these are the clusters with a reasonably high SRA anomaly score. As a result, I do further analysis on each cluster, focusing on the one with the greatest fraud percentage (Anamika et al., 2018).

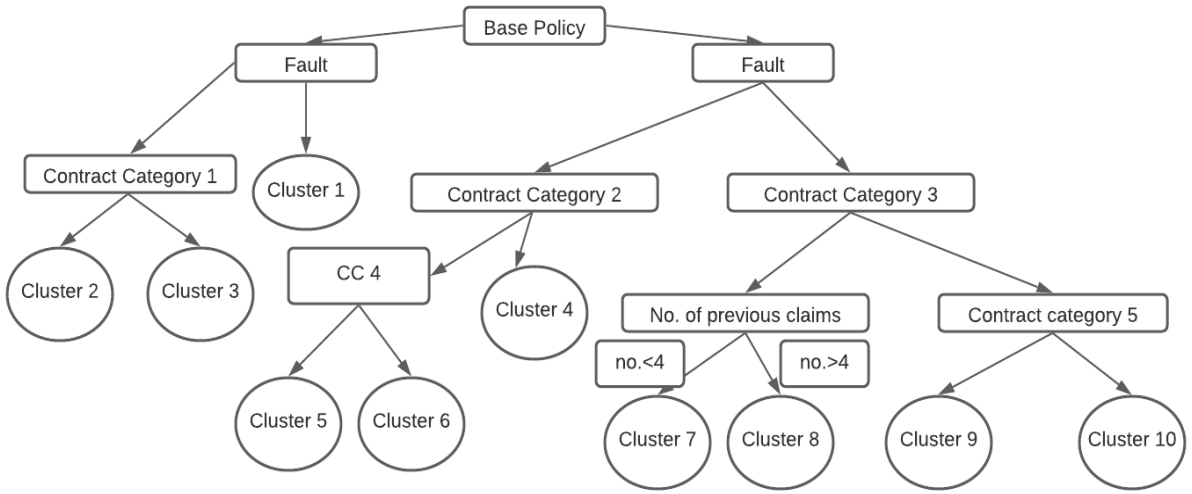

Figure 12: Clusters Decision Tree

A conventional CART (classification and regression tree) is constructed to establish the decision rule for each cluster. The amount of data points at a leaf node must not be fewer than 15 in order to acquire further information. Our hand identification of eigenvectors 1, 2, and 3 serves as our training labels for CART. Next use the whole data set to train CART on these eigenvectors and their labels. The calculated CART tree is shown in Figure 10. From Figure 12, I deduce the following principles for the clusters 4, 6, and 7, to which SRA has awarded high rankings (Singh & S, 2018):

a) Claims with a high fraud ratio are those in which the policyholder caused the accident and the collision or all-risks insurance policy covered the damage (cluster 4, 6, & 7).
b) The equivalent claim goes to the other clusters if the insurance policy covers responsibility and/or the policyholder is not at fault for the accident (third party) (cluster 1, 2, 3, and 5).
c) If the policyholder owns a sports automobile, the appropriate claim falls under cluster 7. If the policyholder has a utility vehicle, the claim falls under cluster 6. Otherwise, the matching claim is classified as a member of cluster 4.

Indeed, these guidelines flag instances of reasonable suspicion. Additionally, as a result of the previous section's supervised random forest training, I determined that the three most critical criteria for classifying fraudulent instances versus lawful cases are the base policy, the automobile type, and the fault. These three characteristics are the ones that are utilized to define the clusters. This research demonstrates that the SRA rating is significant and justified.

CHAPTER SIX: RESULTS AND DISCUSSION

The discussion of the results will present the major findings from the data collected from the literature review, the data analysis and the result chapters. This study analyzed fraud and various causes of fraud in the procurement sector. It also studied the various data mining techniques that can be used in detection of fraud. This project also presents the extent of fraud in the procurement of the companies that participated in the survey. It also presents the analysis of the data mining techniques well suited for procurement fraud detection.

6.1 Data Mining in Procurement Fraud Detection

For the research, I found out the importance of data mining in detecting fraud in procurement. Data analytics reflects on the structuring of the data to be usable and accessible to teams or individuals who require information about procurement in a company. This essentially makes it easy to detect fraud and thus prevents it from happening. On the other hand, data mining is the process by which raw data is turned into information. In this method, the software is used to synthesize patterns of the procurement processes in an organization so that loopholes can be identified and the sourcing of general information.

Data mining involves evaluating and analyzing massive data sets to establish meaningful trends and patterns from which information can be extracted. The process involves data collection, and the next step involves storing and managing data in servers or the cloud. Depending on the intended use, the data is organized, and software is applied to generate actionable data.

6.2 Data Mining Techniques Used in Analysis and Detection of Procurement Fraud

From the analysis of data gathered on data mining techniques, I found that the supervised techniques are efficient when there is a specific value target in the data. You would like to calculate its probability and predict its occurrence. On the other hand, the supervised data mining techniques detect the hidden relation and structure in the data, not focusing on predetermined attributes.

6.3 Fraud Activities in the Procurement Sector

From this research, I identified the following types of procurement fraud that exist in organizations and companies. They include;

a) Employees colluding with suppliers (Velasco & Carpanese & Interian Neto & Ribeiro, 2021).
b) The employees are setting up fraudulent companies to award themselves tenders.
c) Conflict of interest, occurring when someone controls the procurement process.

From our findings, these frauds happen mainly because of the reasons the that are listed below;

a) ***Rationalization:*** This is when the procurement officer tries to justify or explain their actions or thoughts. They do fraud because they are convinced that it is a one-time thing and it will never recur.
b) ***Perceive Opportunity:*** this is when the person identifies a hole in the company's procurement process and wants to use it to benefit themselves instead of covering the holes. In this situation, when the person committing fraud has a good understanding of the controls well or a total lack of controls on the procurement process, they have a wide opportunity for fraud.
c) ***Financial needs:*** This is when the procurement person or company commits fraud because of the need to retain some percentage of the money to use the money to

pay debt, gamble, or any other short-term needs and luxurious needs such as spending on vacations or buying expensive electronics.

CHAPTER SEVEN: CONCLUSION AND RECOMMENDATION

7.1 Recommendation

The use of data mining techniques and control systems is efficient in detecting fraud. Several techniques are discussed herein that are efficient for fraud detection(Khan,2020). From this study, I propose that most organizations implement data mining techniques to fight fraud. With technology, it will be easier to spot fraud in the procurement process before it even happens.

7.2 Summary of the Study

From the research study, I found out that most organizations are dealing with fraud issues in their procurement process. Most of the organizations are losing a lot of resources in fraud during procurement. In every company that participated in our survey, procurement fraud prevention is one of the major priorities. They agree that systematic and robust procurement will help prevent procurement fraud and attract external funding in the business. The procurement can happen in any part of the procurement process. However, auditing has to adapt to the growing amounts of data caused by digital transformation; to adapt to these vast growing demands, there is a need to implement machine learning and data mining, and analytic techniques into auditing. From the analysis of various reasons for procurement fraud based on our survey, I found out that the main reasons are classified into these categories;

a) **Financial needs:** This is when the procurement person or company commits fraud because of the need to retain some percentage of the money to use the money to pay debt, gamble, or any other short-term needs and luxurious needs such as spending on vacations or buying expensive electronics.

b) **Perceive Opportunity:** This is when the person identifies a hole in the company's procurement process and wants to use it to benefit themselves instead of covering the holes. In this situation, when the person committing fraud has a good understanding of the controls well or a total lack of controls on the procurement process, they have a vast opportunity for fraud.

c) **Rationalization:** This is reported when the procurement officer tries to justify or explain their actions or thoughts. They do fraud because they are convinced that it is a one-time thing and it will never recur.

7.3 Recommendation of Future Work

There is always a need to have dynamic systems within our work environment, which is because there is always an arising problem that needs new solutions. This is the same with procurement fraud detection and mitigation. The fraudsters will keep inventing new ways to commit fraud, and scientists have to develop new ways of detecting fraud plans way before they happen. There is a need to look into advancing the data mining techniques to be compatible with machine learning and AI to improve the efficiency and reliability of the detection and control systems. In conclusion, I recommend more research into this study and other similar studies that will help detect fraud.

Conclusion

Fraud detection is a field that requires dynamic research and periodical upgrades and innovations because fraudsters are many and skilled; they consistently devise new ways to perform fraud in a less detectable way. The use of data mining techniques will help discover entirely invisible patterns and alert the fraudsters.

Fraud detection in the procurement sector of most companies in developed and developing countries is needed because there is the inability to find out fraudulent activities when the fraudsters always discover new ways of committing fraud. Fraud detections using advanced technology are needed to protect the procurement sectors in UAE and globally from losing a lot of money and, in return, affect the resources acquirement and performance of the company. Implementing a method of predicting fraud would be an outstanding achievement to the organization's procurement process. Many organizations will be channeling their funds towards acquiring a mechanism to detect and report fraud before they happen. The research from this study showed a need to invest in a system that investigates and detects any abnormal behavior in some of the procurement processes.

REFERENCES

1. Anamika, G., Mayuri, K., Kharthik, R. K., & Ronnie, C. D. (2018). Analyzing the performance of Various Fraud Detection Techniques. *International Journal of Security and Its Applications*. http://dx.doi.org/10.14257/ijsia.2018.12.5.03
2. Haofan, Z. (2016, March). Auto Insurance Fraud Detection Using Unsupervised Spectral Ranking for Anomaly. *The Journal of Finance and Data Science*. 10.1016/j.jfds.2016.03.001
3. S, B. M., D, R. S., & V, V. (2018). Impact of Gradient Ascent and Boosting Algorithm in Classification. *International Journal of Intelligent Engineering and Systems*. 10.22266/ijies2018.0228.05, vol. 11, no.
4. Singh, R., & S, .. M. (2018). Fitting a Neural Network Classification Model in MATLAB and R for Tweeter Data set. *Proceedings of International Conference on Recent Advancement on Computer and Communication*, 10-19.
5. Velasco, R.B., Carpanese, I., Interian, R., Paulo Neto, O.C. and Ribeiro, C.C., 2021. A decision support system for fraud detection in public procurement. International Transactions in Operational Research, 28(1), pp.27-47.
6. V, K. P., S, B. M., & N, I. (2017). Recommendation Engine for Predicting Best Rated Movies. *International Journal of Advanced Science and Technology,*. http://dx.doi.org/10.14257/ijast.2018.110.07
7. Westerski, A., Kanagasabai, R., Shaham, E., Narayanan, A., Wong, J. and Singh, M., 2021. Explainable anomaly detection for procurement fraud identification—lessons from practical deployments. International Transactions in Operational Research, 28(6), pp.3276-3302.
8. Omar, B. and Alturki, A., 2020. A systematic literature review of fraud detection metrics in business processes. IEEE Access, 8, pp.26893-26903.

9. Nonnenmacher, J. and Marx Gómez, J., 2021. Unsupervised anomaly detection for internal auditing: Literature review and research agenda
10. ProcureDesk (2020): Procurement Fraud Definition and Prevention. Available at: https://www.procuredesk.com/procurement-fraud/.
11. Credit Card Fraud (2021 September 30). In Wikipedia. Available at: https://en.wikipedia.org/wiki/Credit_card_fraud.
12. Hussain, S., Atallah, R., Kamsin, A., & Hazarika, J. (2018, April). Classification, clustering, and association rule mining in educational datasets using data mining tools: A case study. In *Computer Science On-line Conference* (pp. 196-211). Springer, Cham.
13. Maia, J., Junior, C. A. S., Guimarães, F. G., de Castro, C. L., Lemos, A. P., Galindo, J. C. F., & Cohen, M. W. (2020). Evolving clustering algorithm based on mixture of typicalities for stream data mining. *Future Generation Computer Systems*, *106*, 672-684.
14. Rohan Ahmed (2018, May 20). *Data mining techniques in financial fraud detection. Grin Verlag (May 20, 2018).*
15. *Diwakar, T & Damodar, R. E& Bhawana, N (2019)*: *Fraud Detection using Data Mining Techniques*. LAP LAMBERT Academic Publishing (March 11, 2019)
16. Maia, J., Junior, C. A. S., Guimarães, F. G., de Castro, C. L., Lemos, A. P., Galindo, J. C. F., & Cohen, M. W. (2020). Evolving clustering algorithm based on mixture of typicalities for stream data mining. *Future Generation Computer Systems*, *106*, 672-684.
17. Yogita Goyal, Anand Sharma (2020). *Credit Card Fraud Detection and Analysis Through Machine Learning*. Available at: *https://books.google.co.ke/books/about/Credit_Card_Fraud_Detection_and_Analysis.html?id=dLa1zQEACAAJ&source=kp_book_description&redir_esc=y* .

18. Al Hosani, N. M. (2020). *Corruption in Construction Industry and Mitigation Actions in the UAE* (Doctoral dissertation, The British University in Dubai (BUiD)).
19. Hussain, S., Atallah, R., Kamsin, A., & Hazarika, J. (2018, April). Classification, clustering and association rule mining in educational datasets using data mining tools: A case study. In *Computer Science On-line Conference* (pp. 196-211). Springer, Cham.
20. Rahul, G.& Amit, K. & Habeebullah, H. (2020). *Review on Credit Card Fraud Detection using Data Mining Classification Techniques & Machine Learning Algorithms.* Available at: https://www.researchgate.net/publication/349624929_Review_on_Credit_Card_Fraud_Detection_using_Data_Mining_Classification_Techniques_Machine_Learning_Algorithms
21. Rahul, R. (1618). TEM Journal; Novi Pazar Vol. 9, Iss. 4, (Nov 2020): Unsupervised Data Mining with K-Medoids Method in Mapping Areas of Student and Teacher Ratio in Indonesia.
22. Itoo, F & Singh, S. (2021). Comparison *and analysis of logistic regression, Naïve Bayes and KNN machine learning algorithms for credit card fraud detection* - International Journal of Information Technology – Springer.
23. Hardinata, D., & Kamaludin, S. W. (2021). Practical Perspective of E-Procurement And Fraud Detection on Public Sector Organization. *Solid State Technology*, *64*(1), 3940-3959.
24. Westerski, A., Kanagasabai, R., Shaham, E., Narayanan, A., Wong, J., & Singh, M. (2021). Explainable anomaly detection for procurement fraud identification—lessons from practical deployments. *International Transactions in Operational Research*, *28*(6), 3276-3302.
25. Carneiro, D., Veloso, P., Ventura, A., Palumbo, G., & Costa, J. (2020, November). Network Analysis for Fraud Detection in Portuguese Public Procurement. In *International Conference on Intelligent Data Engineering and Automated Learning* (pp. 390-401). Springer, Cham.

26. Verma, S., & Rattan, P. INTRODUCTION TO DATA MINING TOOLS AND TECHNIQUES & APPLICATIONS: A REVIEW. *in Business*, 57.
27. Sánchez-Aguayo, M., Urquiza-Aguiar, L., & Estrada-Jiménez, J. (2021). Fraud Detection Using the Fraud Triangle Theory and Data Mining Techniques: A Literature Review. *Computers*, *10*(10), 121.
28. Eswaran, M., Deepa, S., Hamsa Nandini, S., & Ojha, S (2020credic). Identification Of Credit Card Fraud Detection Using Decision Tree And Random Forest Algorithm.
29. Yadav, A. K. S., & Sora, M. (2021). Financial Fraud Detection Using Deep Learning Approach. *Design Engineering*, 6254-6267.
30. Khan, S., & Farooqui, Z. A Review of Effective Intrusion Detection Systems using Data Mining Techniques.
31. D'Allerto, R., & Raggi, M. (2021). From collection to integration: Non-parametric Statistical Matching between primary and secondary farm data. *Statistical Journal of the IAOS*, (Preprint), 1-11.
32. Kalu, A. O. U., Unachukwu, L. C., & Ibiam, O. (2019). Accessing secondary data: A literature review.
33. Johnston, M. P. (2017). Secondary data analysis: A method of which the time has come. *Qualitative and quantitative methods in libraries*, *3*(3), 619-626.
34. Zangirolami-Raimundo, J., Echeimberg, J. D. O., & Leone, C. (2018). Research methodology topics: Cross-sectional studies. *Journal of Human Growth and Development*, *28*(3), 356-360.

35. Peling, I. B. A., Arnawan, I. N., Arthawan, I. P. A., & Janardana, I. G. N. (2017). Implementation of Data Mining To Predict Period of Students Study Using Naive Bayes Algorithm. *Int. J. Eng. Emerg. Technol*, *2*(1), 53.

36. Kiran, S., Guru, J., Kumar, R., Kumar, N., Katariya, D., & Sharma, M. (2018). Credit card fraud detection using Naïve Bayes model based and KNN classifier. *International Journal of Advance Research, Ideas and Innovations in Technology*, *4*(3).

37. Campus, K. (2018). Credit card fraud detection using machine learning models and collating machine learning models. *International Journal of Pure and Applied Mathematics*, *118*(20), 825-838.

38. Khare, N., & Viswanathan, P. (2020). Decision Tree-Based Fraud Detection Mechanism by Analyzing Uncertain Data in Banking System. In *Emerging Research in Data Engineering Systems and Computer Communications* (pp. 79-90). Springer, Singapore.

39. John, H., & Naaz, S. (2019). Credit card fraud detection using local outlier factor and isolation forest. *Int. J. Comput. Sci. Eng*, *7*(4), 1060-1064.

40. John, H., & Naaz, S. (2019). Credit card fraud detection using local outlier factor and isolation forest. *Int. J. Comput. Sci. Eng*, *7*(4), 1060-1064.

41. Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint* (Vol. 48). Cambridge University Press.

42. Harding, J. (2018). *Qualitative data analysis: From start to finish*. Sage.

43. Irawan, Y. (2019). Implementation Of Data Mining For Determining Majors Using K-Means Algorithm In Students Of SMA Negeri 1 Pangkalan Kerinci. *Journal of Applied Engineering and Technological Science (JAETS)*, *1*(1), 17-29.

44. Nayak, J. K., & Singh, P. (2021). *Fundamentals of Research Methodology Problems and Prospects*. SSDN Publishers & Distributors.