5-6-2021

Data Analytics Methods in Preventing Smuggling Drugs

Hessa Abdulla Almarri
haa3448@rit.edu

Data Analytics Methods in Preventing Smuggling Drugs

by

Hessa Abdulla Almarri

A Capstone Submitted in Partial Fulfilment of the Requirements for the Degree of Master of Science in Professional Studies:

Data Analytics

Department of Graduate Programs & Research

Rochester Institute of Technology

RIT Dubai

May 6, 2021

RIT

Master of Science in Professional Studies:

Data Analytics

Graduate Capstone Approval

Student Name**: Hessa Abdulla Almarri**

Graduate Capstone Title**: Data Analytics Methods in Preventing Smuggling Drugs**

Graduate Capstone Committee:

Name: Dr. Sanjay Modak **Date:**

Chair of committee

Name: Dr. Ioannis Karamitsos **Date:**

Member of committee

Acknowledgements

This work would not have been possible without the Rochester Institute of Technology (RIT) in Dubai support. I am grateful to all of the instructors I have had the pleasure to be their student during the Data Analytics program – Cohort 4. In the first semester of the program, the whole world faced the Covid-19 pandemic, and RIT made huge efforts to make virtual learning easy for us. Also, I would especially like to thank Dr. Ioannis Karamitsos for all his support as my teacher and mentor. He has taught me more than I could ever give him credit for here. I would also like to express my deep gratitude to Professor Dr. Sanjay Modak - chair of the committee, for all his support.

Lastly, I would Like to thank my workplace for encouraging me to pursue my master's degree in the data analytics field and providing all the support and needed dataset to accomplish my capstone project.

Table of Contents

TABLE of FIGURES

Key words:

- Illicit
- Drugs
- Terrorist
- Smuggling
- Customs
- Unsupervised approach
- CRISP-DM
- Tableau
- RStudio
- R
- Clustering
- Country of origin
- Concealment Methods
- Drug type
- K-mean
- k-medoid (PAM)
- Hierarchal Clustering (Dendrogram)

Abstract

For the final requirements of (MS) of Data Analytics, we have to work on a capstone project as a graduate student. In the capstone project, we have to implement the data mining techniques we have learned during the program. This capstone project focuses on illicit drugs smuggling, where drugs have a massive negative effect on the countries and individuals. Data mining techniques have been applied to the drug smuggling dataset that has been captured worldwide. The data mining approach used in this capstone project is an unsupervised approach focusing on clustering. Three types of clustering models have been used: k-means, medoid means, hierarchal clustering, and the three models have similar results.

Chapter 1: Introduction

Drug trafficking is a global illicit trade, and it is known that the markets for illegal drugs, primarily cocaine, heroin, and cannabis. These drugs have a massive revenue for terrorist groups due to the high cost and the increasing demand, and by the end, this illicit trade supports terrorist operations worldwide. The countries' customs authorities have many efforts in combating drug smuggling. Still, the smugglers are always one step ahead of the customs authorities in smuggling drugs since it is their work, and it is natural for them to innovate in their work and find new smuggling methods. This report aims to research how illicit drugs move around the world and the methods of smuggling. This research motivation is to conduct a detailed analysis of the interdependence between smuggling illicit drugs and the determinants: the countries that smuggle the illicit drugs, the types of smuggled drugs, the price of the smuggled drugs, and more.

1.1 Background of the Problem

Illicit drug trafficking is one of the most dangerous and harmful trafficking. Illicit drug trafficking has the same adverse effects on every country. The first effect that is known is that drug trafficking is considered one of the most common forms of terrorist groups rely on to finance their organizations and activities. Furthermore, the illicit narcotics trade imposes massive health, social, and enforcement costs on society. According to the National Institute on Drug Abuse, In 2018, more than 67,300 Americans passed away because of overdose, including illicit drugs and prescription opioids (Basu, G., n.d.). The international customs duty is to manage two parallel mandates. The first one is, dealing with the effective facilitation of legitimate and legal trade flows of goods, services, people, and more. The second parallel mandate manages to seize illicit commodities, services, and people going in and out of the country. The global customs authorities conduct their own operations to seize the illicit trades, using the advanced technology they own through intelligence information, international cooperation, and more. However, the smugglers will always be one step ahead of the customs. They will always find a new way to smuggle because it is their work and source of living where the customs authorities will still need to update their techniques seizing illicit trade and minimize the effects of smuggling.

1.2 Project Definition and Goals

The goal of this project is to apply data mining techqniques that has been learnt in Data Analytics program in the past 18 months. By having a better understanding of the drugs smuggling routes. For example, what are the most countries that drugs are shipped from, the most types of smuggled drugs, the quantity of smuggled drugs, total seizures, and more. Secondly, the customs should know what are the methods of the illicit drugs is smuggled worldwide, does the methods changes over the years or not?, and what is the common method used based on the type of smuggled drugs. In addition, to have a better understand of the quantity and value of the smuggled drugs for each type. After having a knowledge regarding all the mentioned goals, the customs will have a clear understanding on illicit drugs trafficking how to seize it and feed their risk engine with the required information.

Chapter 2: Literature Review

Drug trafficking remains a far more infamous illegal commodity smuggled by organized crime, and this has gained systematic exposure in recent decades. Three drug trafficking regulation conventions govern a wide variety of drug-related practices, such as the manufacture, sale, and storage of narcotic substances for medicinal and business reasons. The research by Hartmeier (2018) analyzes the economic impacts of drug smuggling in Mexico. When reflecting only on the Mexican economy from either a financial standpoint, it has been shown that drug gangs negatively affect. The persistent atmosphere of crime and poverty has left its mark on several areas, pushing away industry, people, tourists, and foreign direct investment. This action reveals a significant flaw of the drug lords. Rather than performing their activities far from law firms and several Mexicans' everyday lives, they seize over whole communities in aggressive practices and create chaos by engaging in land battles with rivaling rivals.These economic measures also impact society in different ways. The research by Hagan (2016) analyzes the impact of cocaine smuggling on UK society. Cocaine smuggling is a necessary process that influences countries around the world and must be studied and understanding the extent of its effects. The drug trade is a worth billions of dollars market dominated by cartels that organized crime on even a global scale. All attempts to eradicate the enterprise were futile to date (Burns-Edel, 2016). Coca is mainly grown in Southeast Asia, with the majority split from Colombia, Brazil, and Guatemala. These are the primary cocaine manufacturers, where it is mass-produced and then sold across boundaries to the Bahamas, Malaysia, and Europe.

A study was conducted by Wena, C. et al. (2012), regarding identifying smuggling vessels with data mining Technology. The study provided more understanding and exploration of the advantages of data mining techniques of smuggling crimes. The study focuses on smuggling goods through fishing vessels leaving and returning Taiwan’s ports, It is being carried out in order to provide an alternative solution to the issue of detecting smuggling boats around the world. In

addition to improve the efficiency of human inspection. The authors have applied many data mining techniques in their research. The first technique is the logistics regression (LR) model, and it allows to predict the variables either if it is numerical or categorical and this model is rarely used in the smuggling field. In this research the LR model applied to predict fishing vessels that smuggles goods, as smuggling is determined using a type of binary identification string, like smuggling or non-smuggling. Secondly, the authors used Artificial Neural Networks (ANN) in their research, where ANN is also rarely used in studies for predicting and identifying crime or smuggling behavior. In authors' research, they recommended a careful selection of the prediction variables to improve the model's precision. Thirdly, performance measures was applied in their research. Where, sensitivity is the positive precision shown in the inspection results of the smuggling fishing vessel, which it is determined by TP / (TP+FN). Specificity is the precision of the model evaluating the non-smuggling cases, or the negative precisioin displayed in the inspection results of the non-smuggling fishing vessel, which is determined by TN / (FP+TN). Precision is the precision of this model in terms of smuggling detection, is determined by TP / (TP+FP). The used dataset was splitted into training dataset and test dataset. The training dataset is passed into LR and ANN methods to create related models. 70% of the data logs are used as training data, with the remaining 30% served as a test data to calculate the precision and recall of the models. With respect to ANN, the authors used multilayers perceptions (MLPs). For LR, the authors used backwards step-wise method. However, based on the research that the authors are working on, they have found that the accuracy of ANN model is much better than the LR model (Wena, C., et al., 2012.).

Another study was conducted by S. Appavu et al., in 2008 regarding "Data Mining Based Intelligent Analysis of Threatening E-mail". Where, the process followed in this study can be applied on smuggling crimes. The authors used data mining techniques to detect threatening e-mails, and they have used a supervised approach. The authors have divided the e-mails dataset they have into a labeled training set, like e-mails labeled as belonging to different types. Such as, threatening e-mails and normal e-mails. The models used in supervised learning are Decision Trees (DTs), Support Vector Machines (SVM), and Naive Bayes (NB) with AD Infinitum. The result

of the paper is that the authors found using the AD Infinitum classifier can be successfully used to detect threatening e-mails.

Chapter 3: Data Analytics Methodology

Before starting the Data Mining project, we need to go through cross-industry process for data mining (CRISP-DM) framework, which provides a structured approach to planning a data mining project, and CRISP-DM contains six phases as depicted in the following Figure.1

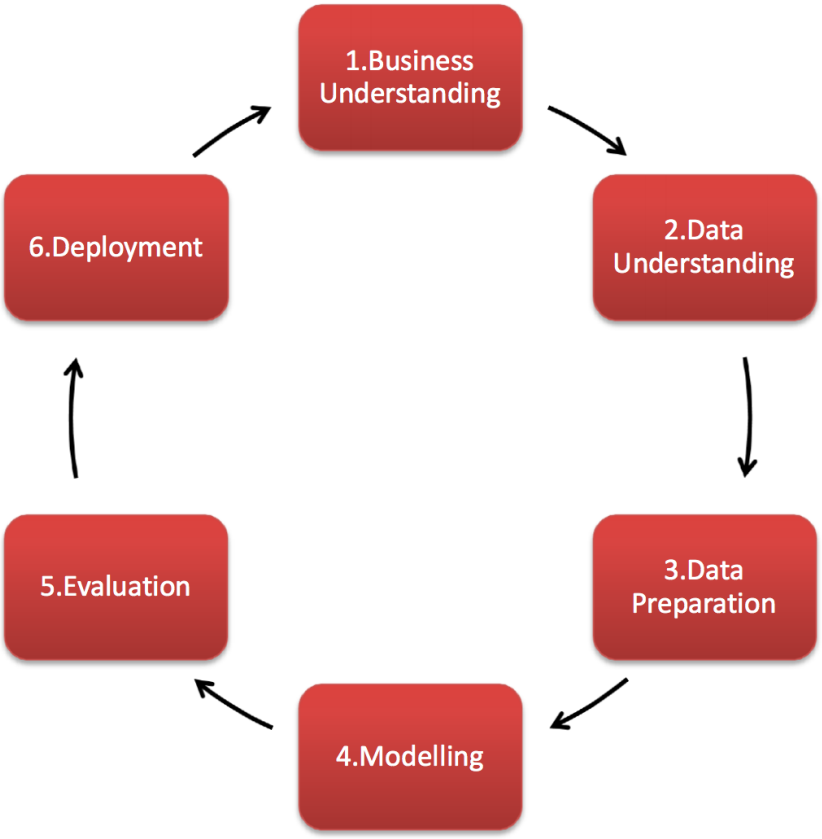

Figure 1: CRISP-DM Methodology (Source: CRISP-DM)

1. **Business understanding:** Before starting with any data mining project, individuals and organizations should understand the problem they are willing to apply data mining to solve. For my capstone project, the requirements for this phase have identified in the statement of

the problem section, background of the problem section, and project definition and goals section.

2. **Data preparation: Gathering data:** I have gathered the required data from my workplace, and I have made sure that it is suitable for the current research I am conducting. After collecting the data, it is essential to discover and understand the dataset by visualizing it. For this step, I have used Tableau to visualize the dataset and understand it. The exploration of the dataset is explained in the analysis section. Before starting the applying of the model, I will manage and clean the dataset. The cleaning step consumes most of the time in the data preparation process, and at the same time, it is also crucial for removing faulty data and filling in gaps.

3. **Modeling:** Then, we have been using many data mining tools and techniques to solve the problems we have given. In this research, I used RStudio and R, for the data mining approach to analyze the dataset, and I have been using unsupervised clustering approach. Where, clustering approach has proven to be an effective method for detecting patterns and structure in both labeled and unlabeled datasets. (Alashwal, et al., 2019)

4. **Results:** After applying any data mining approach and model, we have to write the results of each algorithm outcome which is the core of the project, and any model(s) improvements if needed.
5. **Deployment:** This phase is related to applying the implemented solution in action. For this capstone project, the deployment will be based on the decision of the higher management board in the customs authorities if to proceed with applying it across the local customs or not. However, recommendations and future work will be introduced in this report.

Chapter 4: Data Analysis

4.1 Data Exploration

The dataset used in this capstone project is for the seizures of illicit drugs worldwide for the period 2011 till 2020, and it is provided by the Federal Customs Authority (FCA). The FCA is the entity entitled to deal with customs affairs in the UAE, responsible for customs policymaking and preparing unified legislation for organizing Customs work and anti-customs smuggling and fraud across the seven emirates local customs (“Fca evaluation”, n.d.). However, the dataset used in this work is a rich dataset where it contains more than nine columns and more than 400,000 rows. The following tools were used for this capstone project.

4.1.1 Tableau

Tableau is a visualization tool. It allows individuals and organizations to visualize and manage the data and have a faster insight and make the required decisions (“What is tableau?”, n.d.).

4.1.2 RStudio

RStudio is a graphical programming environment for the R programming language. RStudio features a syntax-highlighting console and editor that facilitates direct code execution, as well as visualization, history, debugging, and workspace management tools. (RStudio Team , 2015). It is used for preparing develop the clustering model and visualize the results.

4.1.3 R

“R is an open source programming language for statistical computing and graphics” (R Development Core Team, 2011)

4.2 Descriptive Statistics and Visualizing the Data

4.3 Data Preparation

4.3.1 Data Preprocessing and Cleansing

The raw dataset cannot be used in the current format in building clustering models. To use the data for clustering, it needs to be prepared to the proper format and cleansed, so all values are numeric, and no null values or anomalies exist as it impacts the clustering heavily. Below are the steps used to prepare the data:

4.3.2 Convert the data to the right format (Crosstab):

What decides the clustering is not the raw data columns, but the occurrence of it values, for example, what is the number of cases where certain type of drugs smuggled in-mail or in-person, another example, what is the common method of drug smuggling based on the country of origin.

Below as an example of converting the data to the right format:

4.3.3 Removing the unusual columns:

Below is the distribution of data based on drug types:

As shown in the figure above, there are types of drugs with very small number of records which cannot contribute effectivity in analyzing and understanding the way of concealment or it is not common and doesn't make a big issue in terms of drug smuggling.

The columns omitted are:

- Concealment types with minimal cases
- Drugs types with minimal cases
- Countries with minimal cases
- Null values
- Unknown values

Next section will show the final datasets used to build the models.

4.4 Data Sample

Based on the above analysis, 25 types of the drugs will be chosen to be clustered against the concealment method.

4.4.1 Drug Type Sample

In the following table.1 is the final sample of data for the model clustering by drug type based on concealment

4.4.2 Country Dataset Sample

In the following Table.2 is the data sample used for clustering by Country of Origin based on concealment method

4.5 Transforming the Data- Standardization (Scaling)

Before passing the data to the model, its values needed to be with the same scale. This process called (Scaling or Standardization). It helps in considering all the values in the dataset regardless of its size compared to other attributes values which prevents dominance of large value while defining the clusters.

Chapter 5: Modeling

After exploring and preparing the data, it is the time to select the right model to resolve the business problem. Below is a summary of Machine Learning different types and algorithms.

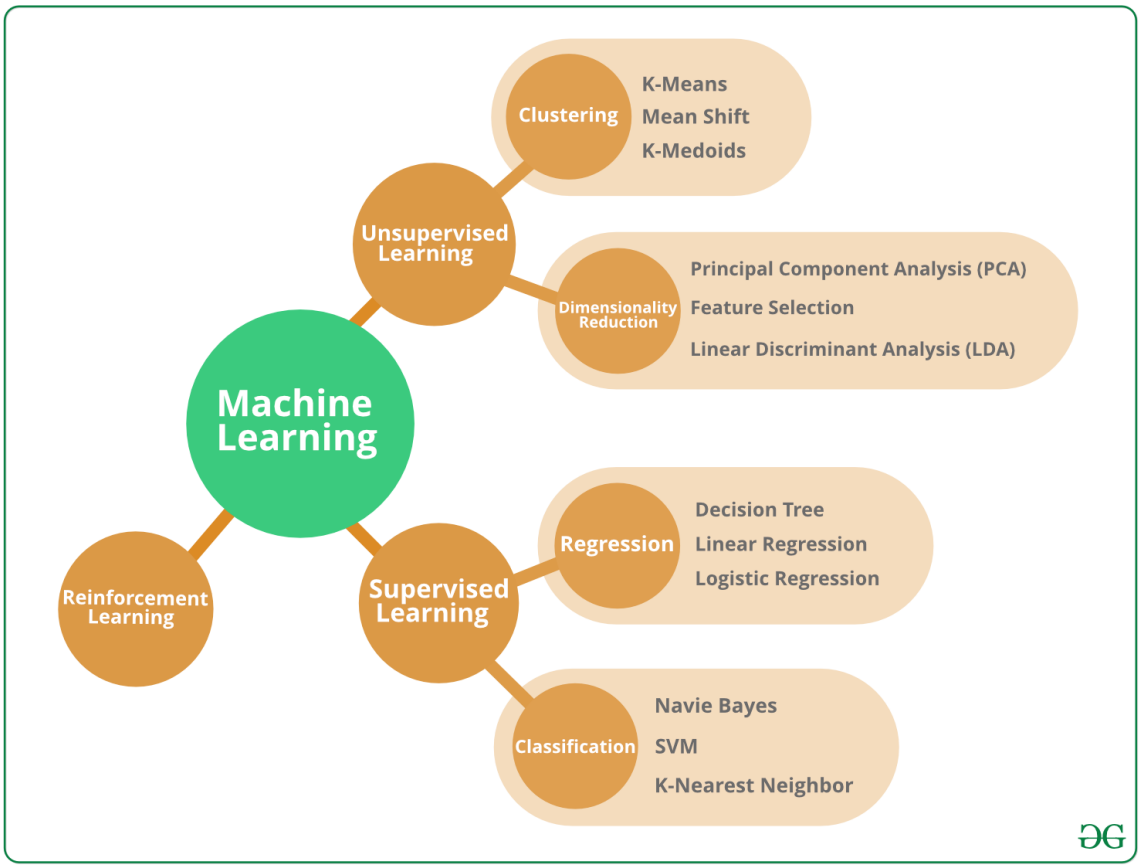

Figure 2: Machine Learning Taxonomy (Source: ("Top 10 algorithms every machine learning engineer should know", 2019))

While the objective of this project is to find the similarities of drug concealment methods for each drug type or country of origin to distribute the resources wisely. This problem falls under Clustering where the model groups the items based on the similarities.

5.1 Model Selection

Many clustering algorithms are used based on the nature of the data and the business problem being resolved. In this project, I will use k-mean, and it is used when we have unlabelled data, which is data without defined classes. This algorithm aims to assign each data point to one

of K groups in the data based on its features (Trevino, 2016). Also, I used k-medoid (PAM). It is related to the k-means algorithms, but K means focuses on minimizing the total squared error, and PAM focuses on minimizing the sum of dissimilarities between points labeled to be in a cluster ("K-means and k-medoids', n.d.). In addition to Hierarchal Clustering (dendrogram), which it attempts to recognize relatively similar groups of variables based on selected characteristics, by using an algorithm that begins with each variable in a divided cluster and combines clusters until only one is left ("Hierarchical cluster analysis", n.d.). The idea of trying more than one algorithm is to find better results as each algorithm has its own advantages and disadvantages.

5.2 Optimal Number of Clusters

The elbow approach examines the percentage of variance explained as a function of the number of clusters; thus, a sufficient number of clusters should be chosen such that adding another cluster does not significantly improve data modeling. As the percentage of total variance by clusters is illustrated against the number of clusters, the first clusters will add a lot of detail, but the marginal benefit will certainly reduce, resulting in an angle in the graph ("Elbow method (clustering)", 2020). The number of clusters is chosen at this stage, hence the "elbow criterion." Since the "elbow" cannot always be clearly defined, this approach is highly subjective and unreliable("Determining the number of clusters in a data set", 2021).

Below the optimal number of clusters based on Elbow Method:

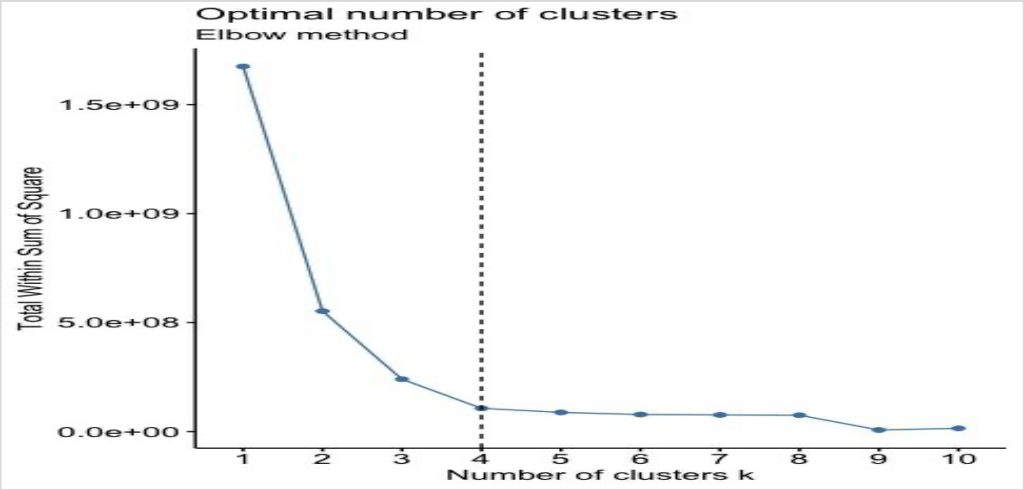

Figure 3: Elbow Method

As shown in the figure above, the optimal number of clusters for each model is 4 clusters.

5.3 Similarity Matrix

Similarity Matrix is used to illustrate the similarity between the data items in terms of distance calculation based on Euclidian method. While it shows the similarity between data points, it doesn't build the cluster which needed to be implemented using clustering algorithm later in this chapter.

The color on the interception point between the item indicate the degree of similarity, the more it is far from red the more the items are similar. Below is similarity matrix between concealment method used for drug types and county of origin.

Chapter 6: Conclusion and Future Work

6.1 Conclusion

To sum up, criminal organizations always find new methods to smuggle illicit drugs. Where these drugs have substantial adverse effects, as explained in the previous sections. The customs authorities and law enforcement authorities should always be updated within the criminal organizations' methods of smuggling and the new routes for smuggling. Different way to stay updated is by using data mining to prevent smuggling illicit drugs. During searching for the right topic for the capstone project, it has been observed that there is a lack of studies regarding using data mining in smuggling crimes. Therefore, I have chosen this topic to implement what I have learned in the past 18 months in the Data Analytics program.

In this project, the data mining approach used is unsupervised, focusing on clustering methods. The clustering methods that have been used to achieve the capstone project objectives are K-means, k-medoid, which is related to k-means, and Hierarchal Clustering. The three algorithms created almost the same clusters. The methods were applied on concealment methods of smuggling by country and by drug type. The results can help the customs authorities and other law enforcement authorities to look for a particular kind of drugs or what to expect to be smuggled from the countries and which countries.

6.2 Future work

I am willing to present my capstone project to the concerned persons in the Federal Customs Authority (FCA) to take the appropriate measures if wanted based on the capstone results. I am also willing to apply data mining methods on other smuggling goods that have negative effects on the country and individuals. Such as cigarettes, dual-use goods, and more.

Bibliography

Appavu, S., Rajaram, R., Muthupandian, M., Athiappan, G., & Kashmeera, K. S. (2009). *Data mining based intelligent analysis of threatening e-mail*. *Knowledge-Based Systems*, *22*(5), 392–393. https://doi.org/10.1016/j.knosys.2009.02.002

Alashwal, H., El Halaby, M., Crouse, J., Abdalla, A., & Moustafa, A. (2019, May 24). *The application of unsupervised clustering methods to Alzheimer's disease*. Retrieved April 01, 2021, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6543980/

Al Arabiya English. (2020, December 25). *Billion-dollar captagon pills seized in Italy smuggled by hezbollah, not isis: Report. Al Arabiya English*. Retrieved April 02, 2021 From https://english.alarabiya.net/News/middle-east/2020/12/25/Billion-dollar-Hezbollah-Captagon-shipment-seized-in-Italy

Basu, G. (n.d.). *Combating illicit trade and transnational smuggling: Key challenges for customs and border control agencies*. *World Customs Journal, 8*, 2nd ser. Retrieved December 1, 2020, from https://worldcustomsjournal.org/Archives/Volume%208%2C%20Number%202%20(Sep%202014)/04%20Basu.pdf

Burns-Edel. (2016). *Environmental Impacts of Illicit Drug Production*. Retrieved from escholarship: https://escholarship.org/uc/item/4w64g29s

Crisp dm methodology. (2020, June 17). Retrieved April 21, 2021, from https://www.sv-europe.com/crisp-dm-methodology/

Determining the number of clusters in a data set. (2021, February 08). Retrieved April 27, 2021, from https://en.wikipedia.org/wiki/Determining_the_number_of_clusters_in_a_data_set#cite_note-1

Elbow method (clustering). (2020, December 11). Retrieved April 27, 2021, from https://en.wikipedia.org/wiki/Elbow_method_(clustering)

Hagan, A. O. (2016). *Cocaine Trafficking and the Social Impact of Cocaine on UK Society*. Retrieved from Forensic Research & Criminology International Journal: https://core.ac.uk/download/pdf/46563043.pdf

Hartmeier, P. (2018). *The Economic Impact of Drug Trafficking in Mexico*. Retrieved from researchgate: https://www.researchgate.net/publication/330468273_The_Economic_Impact_of_Drug_Trafficking_in_Mexico

Hierarchical cluster analysis. (n.d.). Retrieved April 27, 2021, from https://www.ibm.com/docs/en/spss-statistics/27.0.0?topic=features-hierarchical-cluster-analysis

K-means and k-medoids. (n.d.). Retrieved April 26, 2021, from http://www.math.le.ac.uk/people/ag153/homepage/KmeansKmedoids/Kmeans_Kmedoids.html

Rstudio. (n.d.). Retrieved April 27, 2021, from https://www.rstudio.com/products/rstudio/

Top 10 algorithms every machine learning engineer should know. (2019, August 20). Retrieved May 03, 2021,

from https://www.geeksforgeeks.org/top-10-algorithms-every-machine-learning-engineer-should-know/

Trevino, A. (2016, December 6). *Introduction to k-means clustering*. Retrieved April 26, 2021, from https://blogs.oracle.com/ai-and-datascience/post/introduction-to-k-means-clustering

Wena, C., Hsu, P., Wang, C., Wuc, T., & Hsu, M. (2012). *E-government Information Application : Identifying Smuggling Vessels with Data mining Technology.*

What is r? (n.d.). Retrieved April 27, 2021,

from https://www.r-project.org/about.html

What is tableau? (n.d.). Retrieved April 27, 2021,

from https://www.tableau.com/why-tableau/what-is-tableau