K-Means Clustering

with an implementation of Insurance fraud detection

Vincent Chinedu Obigwe
Electronic Engineering
Hochschule Hamm Lippstadt
Hamm, Germany
vincent-chinedu.obigwe@stud.hshl.de

***Abstract*—K-means clustering is a major machine learning data clustering algorithm that comes with the capability of local optimization. It is a numerical unsupervised, repetitive method in which the outcome cannot be determined. It is usually regarded to be fast and effective depending on the usecase it is been applied on. However, this can be argued because there are complaints in terms of performance and the need for improvement. This paper is meant to serve as an introductory guide to the K-means clustering method while pointing out some high and low points. It will be concluded with an implementation of the algorithm in a customer segmentation usecase.**

***Index Terms*—K-means clustering, Machine Learning, Algorithm, customer segmentation**

I. Motivation

In the olden days, humans have applied various kinds of methodologies in performing tasks in a more easy way. The intelligence of the human brain helps in developing tools needed to solve these problems. These machines that were invented by humans have made our lives easy by helping us to do many tasks such as laundry, traveling, etc. Machine learning, therefore, is one of such inventions [9]. In intelligent application, those days systems that are called intelligent were coded with a bunch of ”if” and ”else”, that helps process the data and make a decision on the output based on the users' input. For example, a spam filter has the function of filtering out spam emails based on some already blacklisted words and marking them as spam. Usually, decisions were made using a bunch of decision processes to filter out email, but as things progress it gets difficult because the spam words change, it is therefore important that a system that learns and updates itself is been used. This does not mean that all systems should be based on machine learning, manually creating a bunch of decisions is good, especially for systems where we have a very good understanding of how we could model the process to generate a good output [10].

A. Machine Learning

Machine learning has served as a solution to life long question of how computers can be built such that they can learn from themselves, based on their experience, and improve their actions. It is considered to be one of the highly technical fields that are at the crossroad between computers and statistics. In the machine learning of today, the changes we now see are majorly caused by the development of new machine learning algorithms and theories, also it can be traced to the mass availability of data online coupled with the low cost of computation. Machine learning has so been accepted that it can be seen in all spheres of life from medicine to commerce, which helps us in making decisions from our past events [7]. Machine learning is an ever-changing scope of computational algorithms that are developed to be human-like in terms of intelligence, in such a way that they can learn from their environment also, this is why they are considered the driver of the era of big data [8].

Since machine learning is all about gaining knowledge from data, its application is limitless. We have seen machine learning applications such as recommendations seen on sites like Spotify to image recognition algorithms used by the majority of the social media companies. Apart from the industrial application of machine learning, it further shapes the way in which data-driven research is done. These days machine learning is not only applied to high-scale applications but also to even small applications [10].

Machine learning does not occur in isolation, there are tools and programming language that powers it. Machine learning is mostly done with three major programming languages Python, Matlab, and R. This is because those languages are considered to be data-intensive meaning they work well with data and there are thousands of open source developers that develop modules that is been used within the software. One such module is the Scikit-learn. Scikit-learn is an open-source machine learning module that contains hundred of machine learning algorithms that can be used in solving both supervised and unsupervised machine learning problems. It is supported majorly by python but can also be used in R and Matlab. Scikit-learn makes it easy for one to use machine learning because of the wide range of support, performance, and documentation available [11].

Machine learning algorithms can mostly be classified as either supervised or unsupervised machine learning, even though some authors introduced reinforcement learning as the third type which learns data and patterns for the purpose of reacting to the environment, we will mostly be dealing with supervised and unsupervised machine learning. The major difference between the two is the presence of labels in the training data [12]. From figure 1 we can see not just the division of machine learning, but also the presence of data and

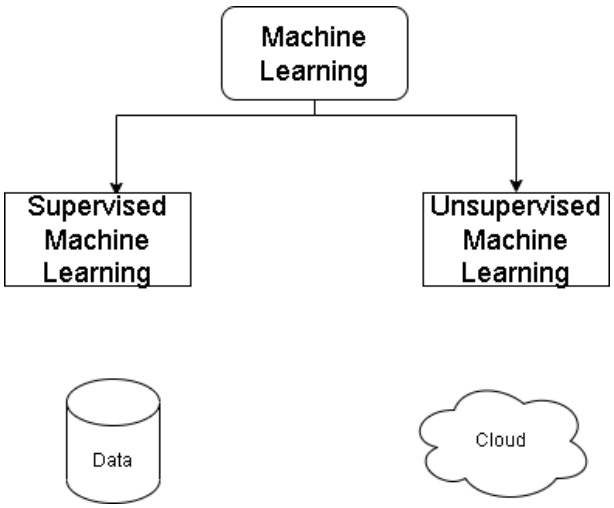

Fig. 1. Classification of Machine Learning

cloud. Data serves as the most important aspect of machine learning because it is the beginning and substance of the process, while the cloud serves as a platform for the machine learning model to be deployed.

B. Supervised Machine Learning

Supervised machine learning is a type of machine learning that has the ability to generate patterns and theories based on some external examples to predict future occurrences [13]. Supervised machine learning can also further be classified as a machine learning that maps a function to input and provides an output based on the previous sample of the input-output provided. This function mapped is inferred from training labeled data. In supervised machine learning, there is a need for external assistance and support. Usually, in supervised machine learning, the input data set is divided into two (train and test) data sets. The training data will contain all the variables including the input and output, while the testing data set does not contain the output to enable one to determine its accuracy. All types of supervised machine learning are built in a way that they learn some patterns in training data and then use the pattern learned to engage in prediction or classification. The figure 2 explains the flow of machine learning from data set to production in a real industrial scenario. Some of the most popular supervised machine learning models include decision trees, Naive Bayes, Support Vector machines, etc [9].

C. Unsupervised Machine Learning

Even though we have a situation where most of the application we have in today's machine learning is composed of supervised machine learning, we still have a high number of data that are unlabelled that we would still need to make sense of. Though unsupervised machine learning can be used on both labeled and unlabelled data, it is mostly remembered when we see unlabelled data [6].

It is referred to as unsupervised learning, due to the fact that unlike in supervised learning, there is no expected output. The

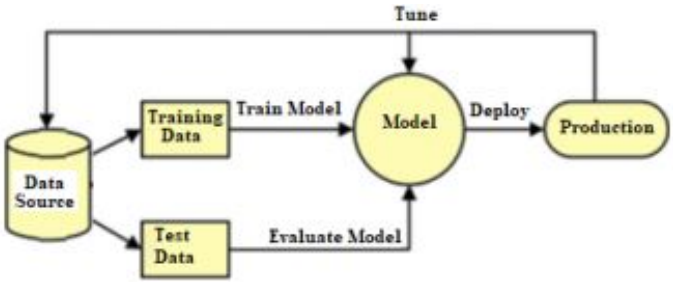

Fig. 2. flow of supervised machine learning [9]

unsupervised machine learning algorithm is left on its own to find out patterns in the data presented to them. After the initial pattern learning, when new data is introduced, it will use the initial pattern to classify the new data. It is mostly used for clustering and feature reduction [9].

1) Feature Reduction: Feature reduction is a feature of unsupervised machine learning where an algorithm helps to make create a new version of data in a way that is easier for either humans or other machine learning algorithms to understand and make use of. a very common application is the reduction of data to two dimensions for visualization [10].

2) Clustering: This is usually used to separate data into several groups of similar items. let's take an example with social media websites. When you upload pictures to social media, to better organize your pictures, the site might decide to cluster your pictures with certain features together, thereby creating several groups from your picture [10].

II. K-MEAN CLUSTERING

K-means clustering is sometimes referred to as the easiest type of unsupervised machine learning in terms of the implementation of algorithms that deals with clustering. The procedure adopted by K-means clustering is such that it deals with the classification of data in various clusters. The major thing is to define K which will serve as the center for each cluster. Care is always when determining the center because it affects the result [9]. In the K-means algorithm, n is the number of available data points divided into K clusters with regard to the similar attributes that the data points possess. The algorithm is very fast and therefore it is one of the most used clustering algorithms. Application of K-means clustering includes vector quantization, Custer analysis, etc [14].

K means clustering works in such a way that after clustering the data, it locates the center of the cluster that is assumed to represent a particular section of the data. This is done by switching between two steps: mapping each data point to the cluster that is nearest to it and set the center of each cluster created as the mean of all the data that belongs to that particular cluster. The algorithm can be considered done when there are no longer changes in the assignment of the clusters. Figure 3 explains how K-mean nearest works on a synthetic data set. In the figure, the clusters are shown as triangles while the circles represent each data point that is part of a cluster. In K-means clustering, one can determine how many clusters they

require from the data set, but care should be taken to avoid over clustering or under clustering a particular cluster because this will determine the insight you might derive from the data [10]. It is important to note that we can assign each data instance to a particular cluster, this is called *hard clustering*, however, when scores are given to the distance between the data point and the center of the cluster, this is called *soft clustering* [6].

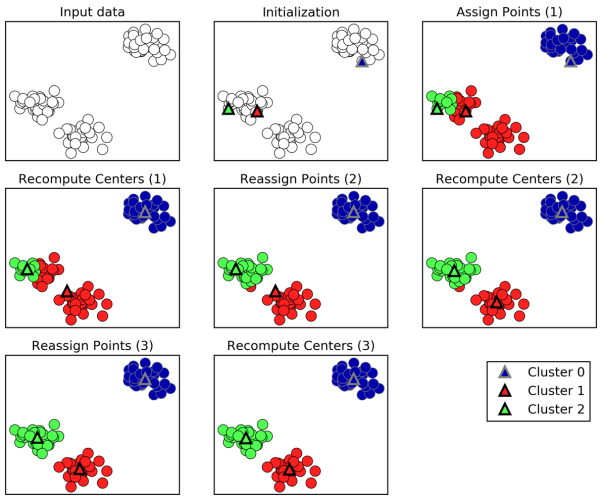

Fig. 3. Three Steps of K-Means algorithm [10]

In figure 3 we indicated that we are searching for three clusters from our data set. The K-means clustering algorithm starts by first initializing three data points randomly to serve as the center of the three clusters we requested, this can be seen in the initialization box. After the first process, the algorithm now assigns each data point to the closest cluster, this can be seen in the assigned point (1) box. Proceeding this step, the cluster center point is recalculated using the mean of the data points assigned, the process does not stop until after two or more iterations when the center of the cluster is no longer changing.

Figure 4 shows the boundaries of the cluster center that we just created. It is quite noted able that applying K-means clustering to scikit-learn is very direct and does not involve many complications. The documentation and supports are fully available online.

To calculate the distance of each data point to the center of the cluster this formula is used:

$$\mathrm{D}(\mathrm{X}_p, \mathrm{C}_j) = \sqrt{\sum_{i=1}^{d} (Xpi - Cji)^2}$$

Xp means the pth data vector, j means the center of the cluster j, while d represents the number of features in each cluster [14].

while to recalculate the new center of the cluster we use:

$$\mathrm{Cj} = \frac{1}{NJ} \sum \forall xp \ \in \ cj \ Xp$$

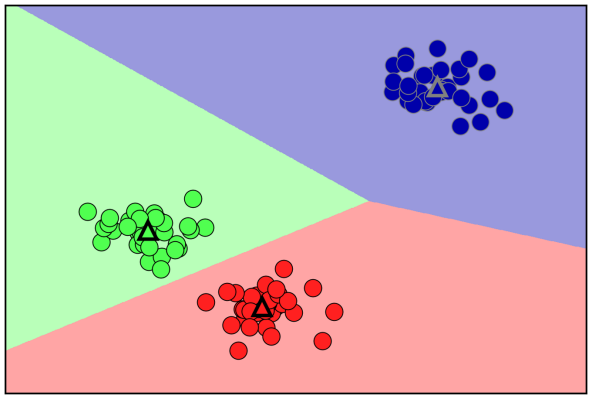

Fig. 4. Diagram indicating the center of each cluster and boundaries [10]

Here NJ means the total number of data points in j while j is the subset of the vector that helps form the cluster Cj [14].

As can be seen from what we have done so far, K-means clustering can be likened to the classification algorithm of machine learning. This is because, at the end of the whole process for both of them, each item gets a label and is grouped into various groups/clusters. The only difference in terms of labeling is that the labels in K-means clustering don't necessarily have a meaning unlike what we have in classification [10].

One of the major shortfalls of K-means clustering is that there is always a problem providing the number of clusters needed. To be able to do this almost accurately, one needs to have a perfect knowledge of clustering.

III. K-Means Clustering Implementation: Customer Segmentation

A. Customer segmentation

customer segmentation possesses the capabilities that are said to be unlimited in terms of helping companies and businesses to learn more impact ways to market their goods and service and also develop new strategies along the line [16]. Customer segmentation is the means by which customers are divided into various sections based on similar features they all possess together. This is so that the company/ business involved can effectively market to each group based on their need. Companies' segments are different when it's business to business or business-to-customer, this is because their customer base is different. In business-to-customer, you consider some personal data such as age, gender, marital status, location, etc while in business-to-customer you consider the industry, no of employee, location, etc [15].

in our implementation, we approached customer segmentation from a business-to-customer (b2c) point of view. This is because we believe that this same approach can be replicated from business to business and effectively achieve the same result.

1) Importance of Customer Segmentation: customer segmentation makes it possible for company marketers to be able

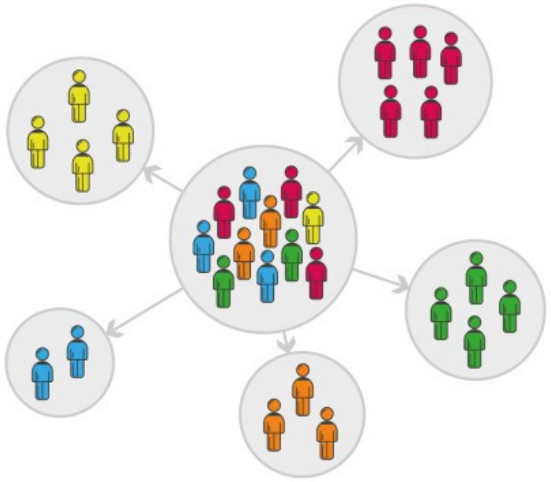

Fig. 5. Diagram Showing a simple example of customer segmentation [17]

to narrow down their work in terms of marketing. This work can be inter-wind with the delivery of communication and development of customer/consumer products. Customer segmentation also becomes more important because of individual differences, we all have different needs and methods we would like to be approached. Some of the ways segmentation helps companies include:

1) Be specific while creating communication for marketing. This is because each section that is been marketed will relate wholly to the message being sent.
2) Help the company select the best communication channel for each segment. Channels such as email, social media, etc. For example, if it is a segment of old people, it may be assumed that social media might not be the best communication channel to be able to reach them.
3) It helps in identifying means by which products and services can be improved while creating opportunities for creating new products and services.
4) Due to the fact that the company understands the customers better, it helps companies establish and maintain a healthy and more profitable relationship with the customers.
5) Helps businesses in testing prices to see the effect or measure customer acceptance.
6) Customer focus: Businesses will now focus more on the profitable customer. This will drive their advert nature and improve their overall profit.

2) How to implement customer segmentation in businesses: The method of customer segmentation is one that requires companies to gather specific data about their customers, run some analysis on it and find patterns that will be used for the purpose of creating segments.

Most of the customer data needed can be gathered from the information available at the point of a customer purchasing goods or services, while some can be taken from customer interaction with our system and others in several other ways. Other methods of gathering information can include:

- Surveys
- Focus groups
- Research
- Face to Face interviews

Having acquired all the required data for customer segmentation, the next thing is to determine the best tools to use [14]

3) Tools for customer segmentation: There are several tools that can be used for customer segmentation, some of them are free while the rest are commercial. One of the free tools frequently used is Google Analytics. Beyond the free tools, one has the capability of creating a custom tool to properly suit the business logic and need. A customer segmentation tool can be created using a simple machine learning algorithm such as K-means clustering. We will show how simple customer segmentation can be created using k-means clustering.

B. Customer Segmentation using K-means clustering

The process of K-means clustering was explained in the section II we will now run a simple implementation of K-means clustering using customer segmentation.

In carrying out customer segmentation, one can either apply the use-case to either business to business or business to consumer. In this instance, our usecase will be applied to business to consumer (B2C). This is important because it will help determine the nature of our data and how we can work around the setup.

1) K-means clustering Data: Several means of gathering data, especially for K-means clustering have been highlighted in the subsection III-A2. Due to the fact that we are approaching a business-to-customer use case, our data will be different from business to business.

	CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	6
3	4	Female	23	16	77
4	5	Female	31	17	40

Fig. 6. Headings of the data used for the machine learning model

As can be seen from figure 6 are data we have to guide our implementations. The goal is to make sense of the data and classify them appropriately. This data is highly limited, it is assumed that in a standard business environment that the data should be rich yet focused to be able to get as much insight as needed.

C. K-means clustering Implementation

To be able to achieve full implementation of K-means clustering there are several processes to it. There are also several programming languages that can be used for machine learning purposes such as Matlab, R, and Python, however, I will be making use of python for my implementation. The

approach I used in building a K-means clustering machine learning model includes:

1) Importing of libraries: Like in most programming languages, importing libraries to be used in building the model is the first thing to be done. The libraries imported include:

- numpy - This is used for data manipulation and handling.
- pandas - This is used also for data works and helps in bringing data into the machine learning model
- matplotlib - This is used for plotting graphs
- seaborn is a data visualization module that is based on matplotlib
- sklearn - This module contains the algorithm used to train our machine learning module. It is one of the most important modules here.

2) Data imports, analysis, and processing: With the help of pandas, our Comma Separated Value (CSV) file that contains our data will be imported into the code. This will then be preceded by basic analysis to determine the shape of our data, check if we have null or missing values, and also choose the appropriate columns that we need for our machine learning model.

3) Choosing the number of clusters: As discussed in the section III-A2 the number of clusters you choose is very important and helps determine how your data will be shaped. What usually happens is that you can go through several iterations with several clusters until you find the one that suits you.

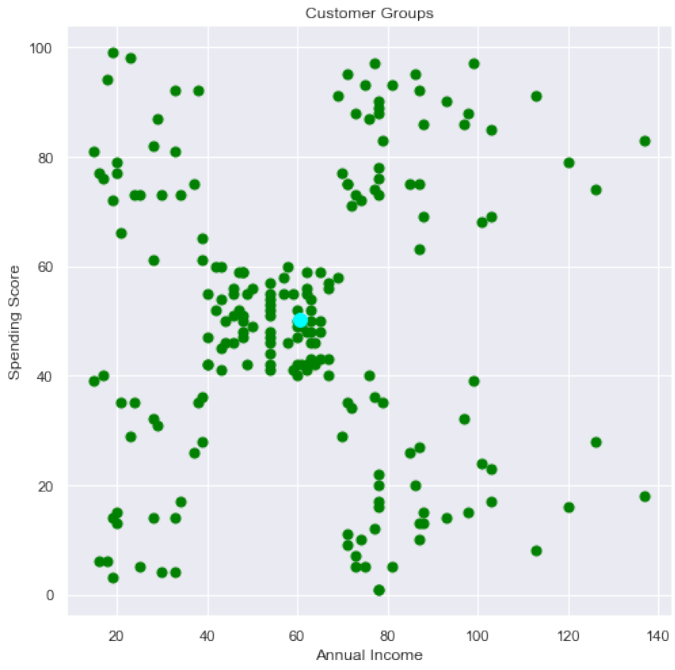

Fig. 7. When cluster is set to 1

As can be seen in figure 7 when the cluster is set to one it doesn't make any sense. As can be observed from our data it is shaped as though it is in five clusters. It is then important that we set our cluster to equal 5 to enable us to capture all the clusters involved in our data.

4) Training of the Model predicting: The K-means clustering model will be trained with the help of KMeans which is part of 'sklearn.cluster'. The module takes an input of the number of clusters (in our case 5), and other basic variables such as 'init', and "random_state". The current training might not take such a long time, but in a standard environment, it is expected that the training will last longer depending on the volume of the data points to be clustered.

After the data has been trained, there is a need to carry out a prediction on the data set to see the labeling of the data points. Such predictions will enable you

```
# Optimum Number of Clusters = 5
kmeans = KMeans(n_clusters=5, init='k-means++', random_state=0)

# return a label for each data point based on their cluster
Y = kmeans.fit_predict(X)

print(Y)
```

✓ 0.2s

```
[4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4
 3 4 3 4 3 4 1 4 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 2 0 2 1 2 0 2 0 2 1 2 0 2 0 2 0 2 0 2 1 2 0 2 0 2
 0 2 0 2 0 2 0 2 0 2 0 2 0 2 0 2 0 2 0 2 0 2 0 2 0 2 0 2 0 2 0 2 0 2
 2 0 2 0 2 0 2 0 2 0 2 0 2 0 2]
```

5 Clusters - 0, 1, 2, 3, 4

Fig. 8. K-means clustering prediction

As can be seen from the prediction in figure 8 it confirms our number of clusters as 5 and also shows the distribution of our data points according to each cluster.

5) Visualizing all the Clusters Created by the Model: After the model predictions, it is supposed that we will visualize all the clusters created before deciding how we could make use of them.

From the figure 9 it can be seen how the clusters were created and the center of each cluster. This will enable the business to be able to create a customer group from it and continue with the rest of the business logic.

6) Deployment: They are several ways that our model can be put into use. One such way is to deploy the model and constantly generate customer groups for whatever business purpose needs to be. The scope of the deployment and usage is not covered by this paper.

IV. Conclusion

K-means clustering is a very interesting aspect of unsupervised machine learning, it helps us to generate insight from our customer data and can be applied in both business to business or business to consumer usecase. This paper gave a brief introduction to machine learning and types before diving deep into K-means clustering and its implementation using customer segmentation.

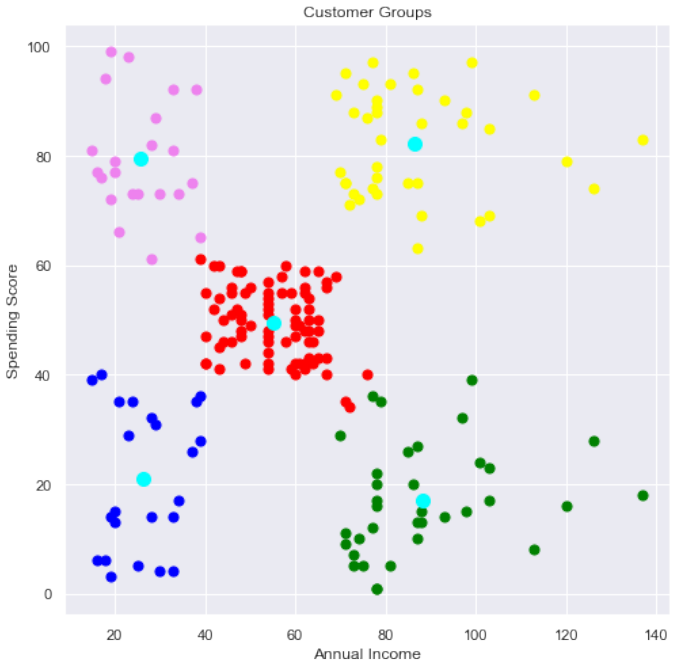

Fig. 9. Visualization of all the cluster created

REFERENCES

[1] Jin X., Han J. (2011) K-Means Clustering. In: Sammut C., Webb G.I. (eds) Encyclopedia of Machine Learning. Springer, Boston, MA.

[2] Na, S., Xumin, L., Yong, G. (2010, April). Research on k-means clustering algorithm: An improved k-means clustering algorithm. In 2010 Third International Symposium on intelligent information technology and security informatics... (pp. 63-67). Ieee.

[3] Lloyd, S. P. (1957). Least squares quantization in PCM. Technical Report RR-5497, Bell Lab, September 1957.

[4] Steinley, D., Brusco, M. J. (2007). Initializing k-means batch clustering: A critical evaluation of several techniques. Journal of Classification, 24(1), 99–121.

[5] MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In L. M. Le Cam J. Neyman (Eds.), Proceedings of the fifth Berkeley symposium on mathematical statistics and probability (Vol. 1, pp. 281–297). California: University of California Press.

[6] Géron, A. (2019). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems. " O'Reilly Media, Inc.".

[7] Jordan, M. I., Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255-260.

[8] El Naqa, I., Murphy, M. J. (2015). What is machine learning?. In machine learning in radiation oncology (pp. 3-11). Springer, Cham.

[9] Mahesh, B. (2020). Machine learning algorithms-a review. International Journal of Science and Research (IJSR).[Internet], 9, 381-386.

[10] Müller, A. C., Guido, S. (2016). Introduction to machine learning with Python: a guide for data scientists. " O'Reilly Media, Inc.".

[11] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.

[12] Berry, M. W., Mohamed, A., Yap, B. W. (Eds.). (2019). Supervised and unsupervised learning for data science. Springer Nature.

[13] Singh, A., Thakur, N., Sharma, A. (2016, March). A review of supervised machine learning algorithms. In 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom) (pp. 1310-1315). Ieee.

[14] Rajeswari, K., Acharya, O., Sharma, M., Kopnar, M., Karandikar, K. (2015, February). Improvement in K-means clustering algorithm using data clustering. In 2015 International Conference on Computing Communication Control and Automation (pp. 367-369). IEEE.

[15] Customer segmentation definition - what is Customer Segmentation. Shopify. (n.d.). Retrieved May 20, 2022, from https://www.shopify.com/encyclopedia/customer-segmentation

[16] Cooil, B., Aksoy, L., Keiningham, T. L. (2008). Approaches to customer segmentation. Journal of Relationship Marketing, 6(3-4), 9-39.

[17] A, P. (2019, May 2). Customer segmentation. Medium. Retrieved May 20, 2022, from https://insights.project-a.com/customer-segmentation-da4967e3a9a4

V. APPENDIX

The code to the implementation https://github.com/vickjoeobi/Customer-Segmentation-Using-K-Means-Clustering

DECLARATION OF ORIGINALITY

I, Vincent Obigwe, herewith declare that I have composed the present paper and work by myself and without the use of any other than the cited sources and aids. Sentences or parts of sentences quoted literally are marked as such; other references with regard to the statement and scope are indicated by full details of the publications concerned. The paper and work in the same or similar form have not been submitted to any examination body and have not been published. This paper was not yet, even in part, used in another examination or as a course performance. I agree that my work may be checked by a plagiarism checker.