Electronic Theses and Dissertations   Graduate Studies, Jack N. Averitt College of

Spring 2019

Digital Forensic Tools & Cloud-Based Machine Learning for Analyzing Crime Data

Majeed Kayode Raji

Follow this and additional works at: https://digitalcommons.georgiasouthern.edu/etd

Part of the Business Analytics Commons, Business Intelligence Commons, Computational Engineering Commons, and the Other Engineering Commons

DIGITAL FORENSIC TOOLS & CLOUD-BASED MACHINE LEARNING FOR ANALYZING

CRIME DATA

by

MAJEED KAYODE RAJI

(Under the Direction of Hayden Wimmer)

ABSTRACT

Digital forensics is a branch of forensic science in which we can recreate past events using forensic tools for a legal measure. Also, the increase in the availability of mobile devices has led to their use in criminal activities. Moreover, the rate at which data is being generated has been on the increase which has led to big data problems. With cloud computing, data can now be stored, processed and analyzed as they are generated. This thesis document consists of three studies related to data analysis. The first study involves analyzing data from an Android smartphone while making a comparison between two forensic tools; Paraben E3: DS and Autopsy. At the end of the study, it was concluded that most of the activities performed on a rooted Android device can be found in its internal memory. In the second study, the Snapchat application was analyzed on a rooted Android device to see how well it handles privacy issues. The result of the study shows that some of the predefined activities performed on the Snapchat application as well as user information can be retrieved using Paraben E3: DS forensic tool. The third study, machine learning services on Microsoft Azure and IBM Watson were used in performing predictive analysis to uncover their performance. At the end of the experiments, the Azure machine learning studio was seen to be more user-friendly and builds models faster compared to the SSPS Modeler in the IBM Watson Studio. This research is important as data needs to be analyzed in order to generate insights that can aid organizations or police departments in making the best decisions when analyzing crime data.

INDEX WORDS: Digital forensics, Autopsy, Paraben E3: DS, Android, Smartphone, Big data, Cloud, Machine learning, Azure, IBM

DIGITAL FORENSIC TOOLS & CLOUD-BASED MACHINE LEARNING FOR ANALYZING

CRIME DATA

by

MAJEED RAJI

B.Eng., Covenant University, Nigeria, 2014

A Thesis Submitted to the Graduate Faculty of Georgia Southern University in Partial Fulfillment of the

Requirements for the Degree

MASTER OF SCIENCE

STATESBORO, GEORGIA

DIGITAL FORENSIC TOOLS & CLOUD-BASED MACHINE LEARNING FOR ANALYZING

CRIME DATA

by

MAJEED RAJI

Major Professor:	Hayden Wimmer
Committee:	Lei Chen
	Weitian Tong

Electronic Version Approved:

May 2019

DEDICATION

I dedicate this to my family for believing in me and their willingness to invest their money on me getting

the best education.

ACKNOWLEDGMENTS

Towards the completion of my thesis, I would like to acknowledge the efforts of my professors, friends, parents and my siblings. I would also like to give special thanks to Dr. Hayden Wimmer and Dr. Lei Chen for their utmost guidance, and Dr. Weitian Tong for his valuable feedback.

TABLE OF CONTENTS

TABLE OF FIGURES

LIST OF TABLES

1   INTRODUCTION

The studies in this thesis show how the application of digital forensics could be of great help during a digital forensic investigation. Digital forensics is a branch of forensic science in which past events can be recreated using digital forensic tools for legal measures. Today, where there are more than a billion Android users all over the world, it shows that its popularity has no equal (Roy, Khanna, & Aneja, 2016). Mobile devices have become increasingly important nowadays due to their capabilities as most users cannot function without their mobile devices because it provides them access to the internet, stores most of their data, allows sending/receiving of calls and messages and so much more. In as much as mobile devices have their advantages, they also have their disadvantages. Criminals tend to use mobile devices in performing various criminal activities. This has led to forensic examiners having to retrieve data and evidence from mobile devices found at crime scenes. Mobile devices nowadays have different browsing applications users can use in gaining access to the internet. The mobile device browsers tend to store the browsing history of the user. This history can be found in the mobile device cache folder for that browsing application. The cache folder of the browser history shows the time and date the user visited a website. Different applications can be installed on mobile devices by the user based on their preference. These applications can be installed and uninstalled based on the interest of the user. More so, the use of social networking applications on a smartphone is quite popular. Recent trends show that inappropriate videos and images are being uploaded on social media by teenagers which on the other hand tend to trigger cyber-bullying and sexting (Burke Winkelman, Oomen-Early, Walker, Chu, & Yick-Flanagan, 2015). Data stored in each application can be of great help during a forensic investigation. This information may include passwords, documents, pictures, messages and much more. The latest technologies in mobile devices are good at storing metadata file of all kinds. The metadata shows information about how files are created and the type of device they were created on. Nowadays mobile phones provide users the opportunity to communicate with others through installed applications. These applications allow users to send and receive SMS, MMS, and emails. These forms of messages can be found in mobile devices and retrieved using forensically accepted tools.

The first study focuses on analyzing data from an Android smartphone while comparing two

forensic tools. The aim of this study is to see if activities performed on an Android smartphone can be retrieved. The two social networking applications focused on in this study are Twitter and the Facebook application. This is implemented by performing several known user activities on the Android device as well as the two social networking applications. The comparison between Autopsy, which is an open source tool, and Paraben E3: DS which is a commercial tool, is performed to know how well they do at retrieving artifacts on the Android smartphone.

The second study focuses on analyzing the Snapchat application on a rooted Android smartphone. Snapchat is a free chatting application that allows users to capture pictures and videos called snaps that self-destruct after a predetermined time (Walker, 2017). The aim of this study is to know the kind of artifacts that can be retrieved to aid forensic investigators during their investigation. This study is implemented by performing a series of user activities on a Snapchat application in a rooted Android smartphone while using the Paraben E3: DS forensic tool.

The third study focuses on analyzing crime data using machine learning service in the cloud. Due to the amount of data being generated daily, there is a need for a better way to store, process and analyze the data. Cloud computing, on the other hand, is a technology that has been of great help at tackling the problems of big data. The aim of this study is to perform machine learning in a cloud environment. This study is implemented by making use of the machine learning service on both Azure and IBM cloud environment to know how well they perform and their ease of use.

2   LITERATURE REVIEW

2.1   Study 1: Analyzing Data from an Android Smartphone while Comparing between Two Forensic

Tools

This study is based on mobile cell phone forensics and it focuses on analyzing data from an Android smartphone while using an open source forensic tool and a commercial forensic tool. With the rate at which technology is evolving in desktops, mobile devices, cloud, notebook computers, and many others, there is a need for new tools and techniques to be created and evaluated in order to sustain the digital forensic field (Wimmer, Chen, & Narock, 2018). This study will expose the strength of the forensic tools to be used to know when each can be applied. This literature review section presents past researches that are relevant to this project and the approaches that were used.

Dogan and Akbal applied digital forensics in analyzing mobile phones. As it is known, digital forensics is the process of collecting, examining, analyzing and reporting of digital evidence without any data damage. Their research work focused on examining and analyzing mobile phones without damaging the data to aid the legal process. Oxygen forensics and MOBILedit were used to perform analyses on an Android phone to ascertain which performs better. After the successful implementation of their research work, it was concluded that MOBILedit tool provided faster analysis compared to Oxygen forensics. However, Oxygen forensics had some advantages over MOBILedit especially in dump analysis, social graph and wireless security protocols (Dogan & Akbal, 2017a).

Thomas et al. performed an analysis of the digital forensic examination of mobile phones. Their research focused on examining the current methods involved in the forensic examination of mobile phones. It also aimed at identifying the areas of mobile phone examination where the current United Kingdom ACPO guidelines and the United States of America NIST guidelines are not clear. The two guidelines were reviewed to demonstrate whether there are specific issues in analyzing a mobile phone that the guidelines did not address. After the successful review of these guidelines, it was found that the issue of remote deletion of data on a phone through an external wireless source was not addressed. It was gathered that both ACPO and NIST guidelines will need to be updated frequently as mobile phones are constantly evolving

(Thomas, Owen, & McPhee, 2010).

Aziz et al. applied mobile device forensics by extracting and analyzing data from an Android-based smartphone. Their research focused on discovering a method of extracting and analyzing data from an Android phone. A case scenario was initiated for a corporate crime to test the acquisition of relevant data. An image copy of the Android phone was created and was examined using the Sleuth Kit Autopsy. SMS, emails, contact list, call logs, images and calendar events data were pulled and analyzed (Aziz, Mokhti, & Nozri, 2015).

Roy et al. research project explored the tools and techniques in Android phone forensics. The research focused on performing a comparative study amongst different forensic tools and techniques. Forensics tools like Volatility, Droid Spotter, XRY, UFED Touch, Oxygen forensic, MOBILedit, Mobile phone examiner plus, Droid watch, NowSecure were compared to know how well they perform. The result was given in a tabular form while comparing characteristics like free or proprietary, the number of devices they support, other platforms they work on apart from Android and so on (Roy et al., 2016).

Awan conducted research on forensic examination of social networking applications on smartphones. The research focused on studying if the internal memory of smartphones stores the activities and actions performed through social networking applications on Apple iPhone, Blackberry phone, Windows phone, and an Android phone. Facebook, Twitter, and LinkedIn were forensically examined and analyzed using the Encase forensic tool. It was found that not even bits of data or evidence could be extracted from Blackberry smartphones. On the other hand, a significant amount of important data was found to be recoverable in the case of Android, Windows and Apple iPhone (Awan, 2015).

Mahajan et al. research addressed the forensic analysis of instant messaging applications on Android phones. Their focus was on identifying the type of data and information that can be found in Android phones' internal memory for instant messengers. File System extractions were used on five Android phones and forensic examination of two commonly used Instant Messengers (WhatsApp and Viber) were performed using the Cellebrite UFED forensic tool. The results showed that a large amount of potential evidence and valuable data can be found on Android phone

internal memory (Mahajan, Dahiya, & Sanghvi, 2013).

Lessard and Kessler's research addressed simplifying cell phone examinations. The research focused on finding a method of extracting and analyzing data from an Android-based smartphone. Forensics tools used in this research were Cellebrite UFED and Access Data FTK. They were used in analyzing the memory image of the HTC Hero Android phone. The result showed that physical extraction is not supported on Cellebrite UFED and it was also concluded that FTK would be the most useful when searching for a specific string text (Lessard & Kessler, 2010).

Lukito et al. compared data acquisition techniques using logical extraction method on unrooted Android device. The research focused on comparing acquisition methods to see how well the methods get data on the phone. The methodology used was comparing the amount of data extracted using different logical extraction techniques such as logical, SDcard Imaging, Android Backup Analysis, and Commercial provider. The commercial tool used in this research is Oxygen forensics. It was discovered that Android Backup Analysis is the best technique for unrooted Android devices without doing a rooting activity (Lukito, Yulianto, & Jadied, 2016).

2.2   Study 2: Digital Forensic Analysis of Snapchat Application on a Rooted Android Smartphone

This study is based on mobile cell phone forensics and it focuses on analyzing data in a Snapchat application on a rooted Android smartphone using Paraben E3: DS forensic tool. This literature review section presents past researches that are closely associated with the project either in the approaches that were used or in the area of focus.

Farnden et al. (Farnden, Martini, & Choo, 2015) discussed the privacy risks in mobile dating. The authors aimed at using forensic techniques on nine well known proximity-based dating apps to determine the kinds of data that can be retrieved from user devices. The result of their experiment showed that several artifacts such as chat messages and user information could be recovered from these apps which raises concern about user privacy (Farnden et al., 2015).

Levinson et al. (Levinson, Stackpole, & Johnson, 2011) researched on third-party application forensics on Apple mobile devices. The authors discovered that third-party applications on Apple mobile devices house information that is relevant to forensic analysts. When users upload messages or media

content using these applications, plain contents are stored in the data partition of the device memory. These artifacts are not limited to time-stamps, geolocational reference, and authentication credentials. More so, similar discoveries were recounted in research work done by Thakur (Thakur, 2013) and Anglano (Anglano, 2014) based on WhatsApp messenger application on an Android device.

In the work by Lee and Chung, third-party instant messengers of Style UI were analyzed using a proposed methodology in digital forensics. In their research work, artifacts from two well-known instant messengers were discovered and analyzed in several ways. The discoveries made in their research could be of great importance during a criminal investigation by providing vital information needed in the court of law (Lee & Chung, 2015).

Husain and Scridbar performed forensic analysis on instant messaging applications on smartphones. Here, the authors studied and reported on three different instant messengers which are AIM, Yahoo Messenger and Google Talk on an Apple iPhone through logical copies of the device. The result of their experiment showed artifacts like AIM unique phrase, screen name, conversation details with timestamp and plaintext password could be retrieved (Husain & Sridhar, 2009).

Chu et al. (Chu, Lo, & Chao, 2013) performed a study on MSN apps and Skype. On rebooting the devices, the authors discovered target strings in the memory of the Android device. More so, while investigating an Android Viber app, Chu et al. (Chu, Yang, Wang, & Park, 2012) were able to retrieve the sent and received messages from the image of the device memory using the number of the sender and the receiver.

Gao and Zhang performed an analysis of WeChat on an iPhone 5 device in order to address a method of analyzing latent evidence. Different types of artifacts were discovered by the authors in different subfolders within the device. These artifacts are not limited to videos, chat messages, audio files, user information and pictures (Gao & Zhang, 2013).

Dezfouli et al. investigated social networking applications on smartphones using Android and iOS platforms. The authors targeted four well-known social networking applications which are Twitter, Google+, Facebook and LinkedIn on both iOS and Android platforms in order to see if the activities performed by users can be retrieved during a forensic investigation. The result of their experiment made it

known that artifacts like uploaded comments, usernames, login information, uploaded posts, personal information, passwords and exchanged messages could be found in the internal memory of those devices. Moreover, recovering these could aid forensic investigators in the court of law (Norouzizadeh Dezfouli, Dehghantanha, Eterovic-Soric, & Choo, 2016).

2.3   Study 3: Cloud-Based Machine Learning of Crime Data

Ahmed et al. proposed a machine learning model that can be used for improving health care services in the cloud environment. The proposed method was aimed at using a parallel particle swarm optimization (PPSO) in a cloud environment to optimize the selections of VMs. More so, they also proposed a new model for predicting kidney disease diagnosis to measure the performance of the VMs. The prediction models used in their research work were neural network and linear regression which were used concurrently in the Azure cloud studio. The linear regression was used to determine the important features that cause chronic kidney disease while the neural network was used to predict the occurrence of chronic kidney disease. The result of their experiment shows that the proposed models outperformed the state-of-the-art models in terms of total execution time, efficiency and accuracy (Abdelaziz, Elhoseny, Salama, & Riad, 2018).

Chen et al. performed a survey on big data where they shed light on the data-intensive applications, challenges, techniques and technologies involved with big data. Big data is defined by the researchers as a collection of large datasets with a great diversity which makes it very difficult to process and analyze using state-of-the-art data processing approaches or tradition processing platforms. Due to the rate at which data is being generated across different platforms and sectors ranging from economic and business activities to social networking activities, data needs to be analyzed in an efficient manner while using appropriate techniques and technologies. In this review work, the researchers basically characterized big data into 3Vs, volume, velocity, and variety. The volume denotes the amount of data that is being generated per second. The velocity indicates the speed at which the data is being generated while variety describes the type of data available for analysis and their sources. Big data can be applied in a plethora of areas like astronomy, medicine, commerce, business, social administration and so on. More so, big data problem lies in data capture, storage, searching, sharing, analysis, and visualization. Some of the known big data techniques

include statistics, machine learning, neural networks, data mining, social network analysis, pattern recognition, signal processing, optimization methods and visualization approaches (Chen & Zhang, 2014).

Qui et al. presented a survey of machine learning for big data processing. In their research work, big data was characterized by 5Vs, volume, velocity, variety, veracity, and value. The value signifies the importance of data being extracted while veracity indicates the trustworthiness or quality of the extracted data. Several machine learning techniques were reviewed as well as the different learning types. The authors also defined machine learning as a field of research that formally focuses on the theory, performance, and properties of learning systems and algorithms. It was made known that machine learning is divided into three categories which are supervised learning, unsupervised learning and reinforcement learning. The supervised learning deals specifically with the training of dataset that have both input and output variables while the unsupervised train datasets without an output variable. Reinforcement learning involves learning from feedbacks received through interactions with an external environment. The researchers also gave an insight into some advanced learning methods which are representation learning, deep learning, distributed and parallel learning, transfer learning, active learning, and kernel-based learning. Also, critical issues of machine learning for big data like large scale of data, different data types, high-speed streaming data, uncertain and incomplete data were discussed (Qiu, Wu, Ding, Xu, & Feng, 2016).

Stephan et al. reviewed models using logistic regression and artificial neural network classification. Their research aimed at summarizing the similarities and differences between the two models and making a comparison with other machine learning algorithms. This research work specifically showcased how a logistic regression and artificial neural network are built, evaluated and the performance indices to be reported (Dreiseitl & Ohno-Machado, 2002).

Viv et al. performed a statistical review on logistic regression. They defined logistic regression as the technique that provides a method for modeling a binary response variable, which takes the value of 1 and 0. In order to model the dependence of a binary response variable on one or more explanatory variables, which can either be continuous or categorical, Logistic regression may be considered (Bewick, Cheek, & Ball, 2005).

According to Michael Cusumano, SaaS and the platform it is based off, Cloud Computing is rapidly

becoming the industry standard in the modern world of today. Although conceived in the late 90s, the innovation it provided through the SaaS business model was not properly recognized until the 2000s, thanks to giants like Salesforce, Google and Amazon. The use of cloud computing platforms is facilitated by API or web services which dictate the protocol to follow or the steps to utilizing the service and as such, moving to cloud-based infrastructure requires one to make API or services that can work with it. The full integration of SaaS and cloud computing over traditional software is a gradual process. Much like its predecessor, it still has its flaws, with developers having to decide whether upgrading previous products to the cloud or developing new products to work with the cloud is the best step towards transitioning to a cloud platform. The customers also have to ask themselves if the move is cost-effective for their business. The cloud computing market itself will remain relevant if the unlimited potential of cloud is utilized to its maximum. With companies like Google and Amazon providing different products under their cloud platform, more variety will exist as opposed to people jumping to just one because it is more popular (Cusumano, 2010).

Ang et al. compared public cloud providers where they discovered several things. In as much that most cloud providers offer pay-as-you-go computing, the different approach being taken in the different cloud providers has led to their problems. In this research work, the researchers developed a CloudCmp approach which systematically compares the cost and performance of cloud providers. With CloudCmp, network services, persistent storage, and elastic computing can be measured across different platforms (Li, Yang, Kandula, & Zhang, 2010).

Cheng et al. used a data mining approach for credit scoring through the application of a support vector machine (SVM) classifier. Due to the rate at which the credit card industry is growing, there is a need to provide a system that automatically calculates the credit score of individuals. Their research aimed at creating an applicant's credit score using two experimental datasets from the UCI machine learning repository. The support vector machine (SVM) classifier was applied as well as neural networks, decision tree, and genetic programming. The result of the experiment showed that SVM classifier performed better than the other classifiers (Huang, Chen, & Wang, 2007).

According to the NIST, cloud computing is defined as a "model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources that can be

rapidly provisioned and released with minimal management effort or service provider interaction." Some notable cloud computing characteristics are the on-demand request, self-service, fast network access, resource pooling, rapid elasticity, measured service. There are three major service models which are software as a service (SaaS), Platform as a Service (PaaS) and Infrastructure as a service (IaaS). The SaaS offers applications for consumers on cloud infrastructure. The PaaS provides platforms for consumers to host their applications while the IaaS provides infrastructure for consumers to store and process their applications in the cloud (Mell & Grance, 2011).

Gai and Li performed a review on cloud computing and its development trends. This review was based on the definitions, characteristics, operations, security management, service governance and development trends of cloud computing. There has been an urgent demand for information sharing, software and resources over the internet due to the development of internet technologies and its demands. By October 2007, cloud computing was introduced to consumers to tackle this urgent demand by combined efforts of I.B.M and Google. More so, cloud computing has been beneficial in ways which include low-cost, high expandability, availability on the network, environmental protection, innovation power, and friendly utilization. Unfortunately, using cloud computing could be risky due to potential risks like a financial burden, security issues, service agreement issues, low controllability and lack of auditing features. Some of the development trends in cloud computing are Content Delivery Network (CDN) and MetaCDN, VM-centric cloud services due to burst resource demands and integration of grids and clouds (Gai & Li, 2012).

3   STUDY 1: ANALYZING DATA FROM AN ANDROID SMARTPHONE WHILE COMPARING

BETWEEN TWO FORENSIC TOOLS

3.1   Introduction

Nowadays with more than a billion Android users all over the world, it is evidence of the unmatched growing popularity of Android devices (Roy et al., 2016). Due to the increase in the availability of these devices, the potential of their use in criminal activities should also be considered. Data residing in smartphones can be exceptionally useful during forensic investigation due to the large storage space available in smartphones. Smartphone data could contain useful information about a criminal such as call history, contact list, text messages, browser history, chat logs, social networking app messages and so on (Thomas et al., 2010). This study is aimed at comparing the results of the analysis of data retrieved from an Android smartphone by using a commercial and an open source forensic tool.

The Paraben E3: DS tool is chosen amongst other available commercial tools due to its data carving ability, app data parsing, password bypass, and unique content analysis. On the other hand, Autopsy is chosen because of its simplicity, timeline analysis, hash filtering ability, file system analysis, keyword searching and moreover, it is free of cost.

3.2   Methodology

This study is focused on analyzing data from an Android phone while comparing two forensics tools. The comparison was conducted using an open source forensic tool and a commercial forensic tool. This is necessary to know which amongst the forensics tools performs the best under the same testing conditions. File System extraction was conducted on the Android phone and forensic examination will be performed. Mobile forensics came into existence due to the high rate at which crimes were being conducted through mobile devices. During the investigation of Android devices, a forensically sound methodology on data extraction, generating a hash value and time stamps are required to prove evidence in the court of law. Therefore, all the experiments in this study were performed using forensically accepted tools under the required condition by the National Institute of Standards and Technology (NIST). This will preserve the

integrity of data on the Android phone so the evidence can be accepted in the court of law. The forensic analysis will be performed on a rooted Android phone using Paraben E3: DS and Autopsy forensic tools.

Paraben E3: DS is a full-featured advanced mobile device forensic extraction and analysis toolkit with over a decade year of development. The paraben E3: DS tool supports password bypassing, logical extractions, file system extraction and physical extractions of a plethora of mobile devices as well as the Internet of Things (IoT) devices. It supports the different operating systems on smartphones available in today's market. It also can image data from cloud accounts like Twitter, Facebook, and many other social networking applications. In addition, it can recover deleted contacts, SMS history, MMS history, call history as well as non-deleted system files. Due to its capabilities, the E3: DS was ranked as the most sort after forensic tool and also recommended to forensic laboratories at the SCMag Best Buy option ("Paraben Corporation,").

An Autopsy forensic tool is an end-to-end open source digital forensics platform. It is built by Basic Technology with most of the features available in commercial forensics tools such as timeless analysis, hash filtering, keyword search, data carving, and web artifacts extraction. It is used by law enforcement, corporate examiners and military to investigate what happens in a computer and a mobile device. It can be used to recover deleted files from a device memory or its memory card. Its parallel way of running background tasks while using multiple cores make it very fast. An Autopsy forensic tool is cost effective as it is free, and it offers the same core features as other forensic tools. Its ease of use is based on its design to be spontaneous thereby making it easy to use ("Autopsy Forensic Tool,").

3.2.1   Testing Environment and Equipment

A standalone workstation with a system feature that accommodates both forensic tools was set up for performing the test examinations. Table 3.2.1 highlights the version of Android device used and the list of software used for the analysis.

No	Product	Version
1	Autopsy	4.5.0
2	Paraben E3 DS	1.4

3	Android Phone	Nexus 6P, 7.1.2
4	Facebook App	145.0.0.37.86
5	Twitter App	7.17.0
6	Google drive	2.7.372.10.40
7	Titanium Backup	8.0.2
8	Root checker	3.1.1
9	Super User	2.79-SR3

Table 3.2.1: Testing Environment

3.2.2   Analysis and Testing

The testing process and analysis encompasses mainly four stages which are:

3.2.2.1   Rooting

Since Android is based on the Linux kernel, gaining root access on Android enables full access to the system directory. In order to gain the root access, the Android phone was rooted and SuperSU was installed during the rooting process. The SuperSU is the gatekeeper of the root access on the Android device. It approves the request made by applications in order to gain root access.

3.2.2.2   Scenario

This stage involves downloading applications, installing applications manually and conducting various activities. The applications downloaded and installed were WhatsApp, Facebook, Twitter, Google drive, Titanium Backup. Common user activities were performed using the phone such as placing phone calls, receiving phone calls and several more. A set of predefined activities were carried out using the feature of each application installed on the Android phone. The activities performed can be seen in Table 3.2.2 below.

Application	Activities Performed
WhatsApp	Instant messaging Sending photos Sending videos Receiving photo Receiving video Receiving audio
Facebook	Login with the test profile Status update Changing the profile picture Chatting messages

	Checking comments on the wall
Twitter	Login with the test profile
	Tweets
	Following
	Watching a video on the timeline
	Viewing pictures on the timeline
Google drive	Storing file on Google drive
Call app	Receiving phone calls
	Placing phone calls
SMS app	Receiving SMS
	Sending SMS

Table 3.2.2: Application and Activities

3.2.2.3   FileSystem Acquisition

This stage is directed at obtaining the logical image of the Android phone internal memory. The Android operating system (OS) does not provide distinctive software for managing and backing up purposes. Nevertheless, several companies have developed their software for obtaining a device backup like Titanium Backup. This backup can be stored on cloud services or a phone's SD card memory. The Titanium backup application was downloaded and installed on the Huawei Nexus 6P phone to obtain the logical image of the phone internal memory. The SD card was initially emptied before the backup process due to the laid down guidelines of digital forensics analysis in order to protect data integrity. Figure 3.1 is a flowchart showing the processes involved in performing a logical acquisition process.
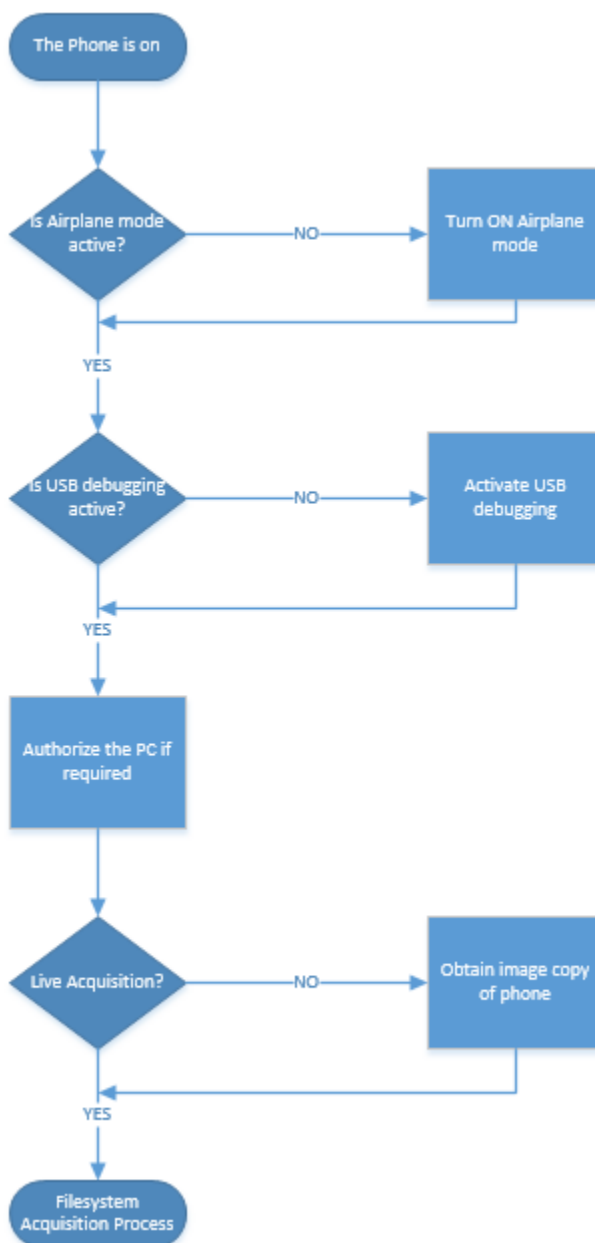
Figure 3.1: FileSystem Acquisition Process

3.2.2.4   Analysis

The logical image obtained through file system acquisition can now be examined using the Autopsy and the Paraben E3: DS forensic tools.

3.3   Results

The artifacts retrieved by using Paraben E3: DS and Autopsy forensic tools are discussed in this section. Figure 3.2 illustrates the relevant extracted information related to this investigation. This includes text messages, Contact list, Facebook artifacts, WhatsApp artifacts, Twitter artifacts, Google drive artifact, and Pictures. The cloud artifacts were extracted using the in-built cloud capture functionality on the Paraben
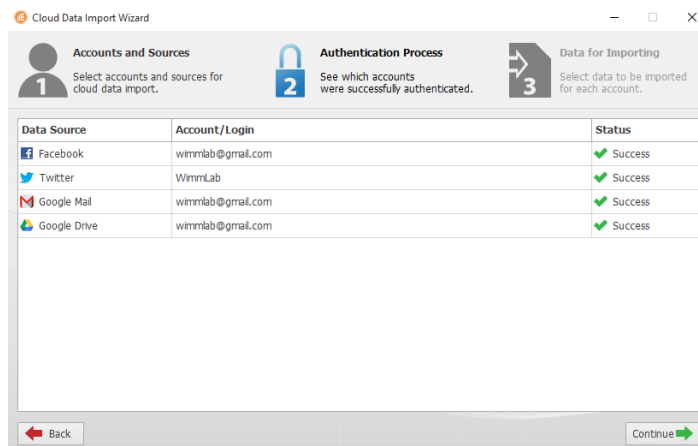
E3: DS.

Figure 3.2: Cloud Import

3.3.1   Facebook Artifacts

Table 3.3.1 below shows the artifacts found due to the activities performed on the Facebook application. Paraben E3: DS was able to retrieve all the activities performed on Facebook application. On the other hand, Autopsy failed at retrieving the required result.

Facebook Artifacts	Paraben E3: DS	Autopsy
Profile name	Yes	No
Profile picture	Yes	No
Status update	Yes	No
Messages	Yes	No
Friends	Yes	No
Picture posted	Yes	No
Profile info	Yes	No

Table 3.3.1: Facebook Artifacts

3.3.2   Twitter Artifacts

Table 3.3.2 below shows the artifacts found due to the activities performed on Twitter. Paraben E3: DS which is a commercial tool is seen to have outperformed Autopsy. The Paraben E3: DS was able to retrieve the activities performed on the user's Twitter account. Figure 3.3 shows the case content on the cloud data using the in-built functionality on the Paraben E3: DS.

Twitter Artifacts	Paraben E3: DS	Autopsy
User handle	Yes	No
User avatar	Yes	No
User location	Yes	No

Direct messages	Yes	No
Followers	Yes	No
Followings	Yes	No
Tweets	Yes	No

Table 3.3.2: Twitter Artifacts
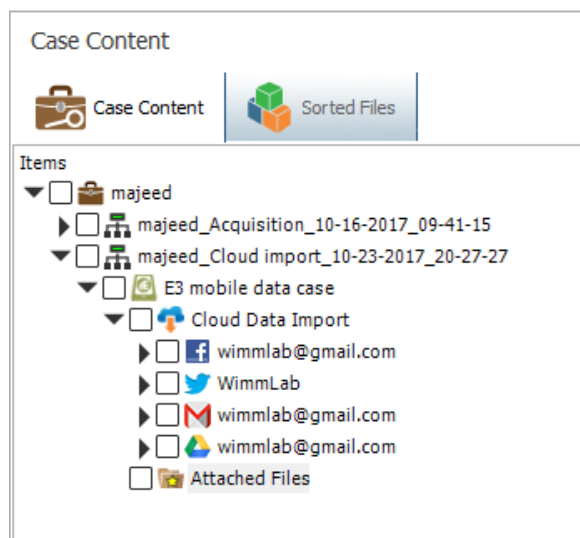
Figure 3.3: Case Content

3.3.3   WhatsApp Artifact

Both Paraben E3: DS and Autopsy forensic tools were able to retrieve the activities performed on WhatsApp, however, Paraben E3: DS was also able to retrieve the time at which the activities were performed. The detailed summary of the result is included in Table 3.3.3 below.

WhatsApp Artifacts	Paraben E3: DS	Autopsy
Audio received	Yes, with a timestamp	Yes, without timestamp
Profile picture	Yes, with a timestamp	Yes, without timestamp
Contacts	Yes, with a timestamp	Yes, without timestamp
Contacts display picture	Yes, with a timestamp	Yes, without timestamp
Video sent	Yes, with a timestamp	Yes, without timestamp
Video received	Yes, with a timestamp	Yes, without timestamp
Sent messages	Yes, with a timestamp	Yes, without timestamp
Received messages	Yes, with a timestamp	Yes, without timestamp

Table 3.3.3: WhatsApp Artifact

3.3.4   Google Drive Artifacts

The data stored in Google drive was also retrieved using the in-built cloud functionality of Paraben E3: DS while Autopsy failed to do so. In addition, it was found that Paraben could not retrieve the activities performed as seen in Table 3.3.4 below.

Google Drive Artifacts	Paraben E3: DS	Autopsy
Shared folder	Yes	No
Google pictures	Yes	No
My drive	Yes	No

Table 3.3.4: Google Drive Artifact

3.3.5 Other Artifacts

The results gathered from the other activities performed on the Huawei Nexus 6P are recorded in Table 3.3.5 below. The call history information was retrieved with timestamp while specifying the call type on Paraben E3: DS. Also, the pictures taken with the Android smartphone were recovered using both forensic tools with the Paraben E3: DS showing the geotag.

Other Artifacts	Paraben E3: DS	Autopsy
Phone Applications (42)	42	41
Call History (14)	Yes, with a timestamp	Yes, without timestamp
Phone Contacts (7)	Yes, with a timestamp	Yes, without timestamp
Call Types (Incoming/Outgoing)	Specified	Not Specified
Call Duration	Yes	Yes
Missed/Received Call	Yes	Yes
Digital Camera Images (5)	Yes, with a timestamp and geotag	Yes, with a timestamp
Screenshots (1)	Yes	Yes
Text Messages (14)	Yes, with a timestamp	Yes, without timestamp

Table 3.3.5: Other Artifacts

3.3.6 Summary

The activities performed on Facebook, Twitter and Google Drive applications were retrieved using Paraben E3: DS while Autopsy failed to do so. Also, the activities performed on the WhatsApp application were retrieved using both Paraben E3: DS and Autopsy forensic tools but Paraben E3: DS surpasses Autopsy with its ability to retrieve the WhatsApp artifact with a timestamp. In addition, other activities such as Call History, Phone Contacts, Call Types (Incoming/Outgoing), Call Duration, Missed/Received Calls, Digital Camera Images (DCIM) and Screenshots were all retrieved using both forensics tools. The commercial tool Paraben E3: DS was able to retrieve the call history, phone contact, and text messages with timestamp while including the geotag of the pictures taken with the Android smartphone camera. Finally, all the application in the Nexus 6P smartphone were retrieved using the Paraben E3: DS tool while the Autopsy failed

to do so.

3.4   Implication for Practice

There are some tools required by digital forensics investigators during their investigation. These tools should allow the investigator to analyze systems in such a way that the results are forensically acceptable. In order to address the issue of evidential integrity, the forensic investigator will work on digital image/copy of the device to be investigated, thereby ensuring the device is not altered in any way (Ayers, Brothers, & Jansen, 2013). Mobile phones found in crime scenes may be found in different states, therefore generating the imaging of the device becomes more difficult. These states include the nascent state, active state, quiescent state, and the semi-active state. The nascent state signifies that the device has no user data (factory fresh). The active state signifies the phone is powered on and specific tasks are being performed on the phone. The quiescent state is when the phone appears to be inactive even though the phone is actually performing functions while maintaining user data such as keeping the time and date accurate (Dogan & Akbal, 2017b). The semi-active state is triggered when the device performs a preset task such as the alarm clock. Therefore, forensic investigators should know the states of mobile phones when they are found in a crime scene. More so, it is required to put the device in an airplane mode in order not to tamper with the state of the mobile phone. This may prevent remote deletion of data from the mobile phone by the offender. It is also advised not to perform live analysis on the evidence so as not to alter its evidential integrity, but a write blocker could be used if it is deemed necessary. These processes must be documented while giving a reason for performing live analysis on the evidence. Additionally, digital forensic investigators should follow the ACPO and NIST guidelines while using forensically accepted tools when trying to analyze a piece of evidence to preserve its integrity in the court of law.

3.5   Conclusion

This study was aimed at analyzing data on Android phone while comparing between a commercial tool Paraben E3: DS and open tool Autopsy. However, all the activity performed on the Android smartphone was retrieved when the commercial tool was used as Autopsy could not do many retrievals. Using the Paraben E3: DS, almost all the activities performed were found with a timestamp, unlike Autopsy. It can be concluded that data residing on Android smartphones can be extracted using the right forensic

tools. These artifacts found can be of great value to forensic investigators during a crime investigation. As the focus in this research is not based on deleted files, retrieving deleted artifacts could be considered in the future work. Also, other social networking applications could be considered besides Twitter and Facebook to see if activities performed on them could be retrieved.

3.6   Future Work

In this study, we examined the performance of Paraben E3: DS and Autopsy forensic tools and the Paraben E3: DS forensic tool was found to be more effective. As the focus in this research is not based on deleted files, retrieving deleted artifacts could be considered in the future work. Also, other social networking applications, like Snapchat, could be considered to determine if activities performed could be retrieved. In the next study, we will address these limitations by performing forensic analysis on the Snapchat application in an Android smartphone.

4   STUDY 2: DIGITAL FORENSIC ANALYSIS OF SNAPCHAT APPLICATION ON A ROOTED

ANDROID SMARTPHONE

4.1   Introduction

The use of social networking applications on a smartphone is quite popular. Snapchat is a free chatting application that allows users to capture pictures and videos called snaps that self-destruct after a predetermined time (Walker, 2017). Recent trends show that inappropriate videos and images are being uploaded on the social media by teenagers, which on the other hand tends to trigger cyber-bullying and sexting (Burke Winkelman et al., 2015). More so, previous researches have proven that most of the activities performed on smartphones can be found in their internal memory which can serve as evidence during a forensic investigation. Some common user activities such as adding friends, sending and receiving messages, pictures and videos were performed on the Snapchat application. This paper is aimed at analyzing artifacts that can be retrieved on Snapchat in a rooted Android smartphone by using Paraben E3: DS forensic tool. The paraben E3: DS tool was selected because it supports a plethora of devices accompanied by its app data parsing ability, keyword search, and cloud access. It also supports all operating systems available on smartphones. At the end of this study, the result showed that some and not all the activities performed on a Snapchat application can be retrieved due to the security measures provided by the social networking application. These retrieved artifacts could be of great help when presented in a court of law as necessary forensic guidelines were complied with to aid its acceptance.

4.2   Methodology

This study work is focused on analyzing artifacts on a Snapchat application on a rooted Android smartphone in order to determine if the activities performed on the Snapchat application can be found in the internal memory of the Android smartphone. Snapchat is a free chatting application that allows a user to capture pictures and videos called snaps that self-destructs after a predetermined time. Recent trends show that inappropriate videos and images are being uploaded on social media by teenagers which on the other hand tend to trigger cyberbullying and sexting.

This section of the study shows the forensically sound methods used to aid the acceptance of the result in the court of law. The tools used in this study is based on the requirements of the National Institute of

Standard and Technology (NIST) (Ayers et al., 2013). This will aid during the process of preserving and acquiring the data stored in the Android device to protect evidential integrity. The forensic tool used in this study is Paraben E3: DS.

4.2.1 Testing Environment and Tools

The tools and software used towards the implementation of this study can be found in Table 4.2.1 below.

No	Product	Version
1	Paraben E3: DS	1.4
2	Google Nexus 6P	Android 7.1.2
3	Snapchat	10.29.0.0
4	Dell Inspiron 13	7000/16G ram

Table 4.2.1: Tools and Version

4.2.2 Analysis and Testing

1. Rooting
2. User Activities
3. File System Acquisition
4. Examination and Analysis

4.2.2.1 Rooting

To gain access to the file system directories in the Android device, the device must be rooted. Rooting the device grants superuser access to the requests made by applications in the device.

4.2.2.2 User Activities

This section deals with the installation of a Snapchat application and how to set up a test account. A problem was encountered when trying to set up the test account as Snapchat does not allow setting up an account on a rooted device. A root switch application is then downloaded and installed to hide root for the Snapchat application. The test account is then created and used in performing some known user activities by exploring the features offered by the Snapchat application using its default settings. The activities performed on the Snapchat application are listed below:

- Setting up the test account

- Receiving an automated snap message from Snapchat

- Adding friends to test account

- Posting stories (pictures/video)

- Placing and receiving audio/video calls

- Sending and receiving snaps

Table 4.2.2 below shows the count of each activity performed on Snapchat.

User		1
Friend		5
Sent	Picture	1
	Video	2
	Message	3
Received	Picture	1
	Video	2
	Message	6
Story	Picture	2
	Video	1
	Message	1
Placed Calls	Audio	2
	Video	1
Received Calls	Audio	1
	Video	2

Table 4.2.2: Count of User Activities

4.2.2.3   Filesystem Acquisition

Android device operating system is built on SQLite database platform. This simply indicates that when data stored in the mobile device database is deleted, it can be retrieved later as they are only marked deleted and not completely overwritten. To acquire data stored in allocated memory space, a live acquisition was conducted on the Android smartphone using the Paraben E3: DS forensic tool. The smartphones airplane mode was activated to disallow the phone from connecting to radio signals. Also, USB debugging was turned on the Android device to allow communication between the forensic tool and the Android

smartphone. The filesystem acquisition took about 75 minutes and was ready for analysis. Figure 4.1 below

shows the filesystem acquisition process.
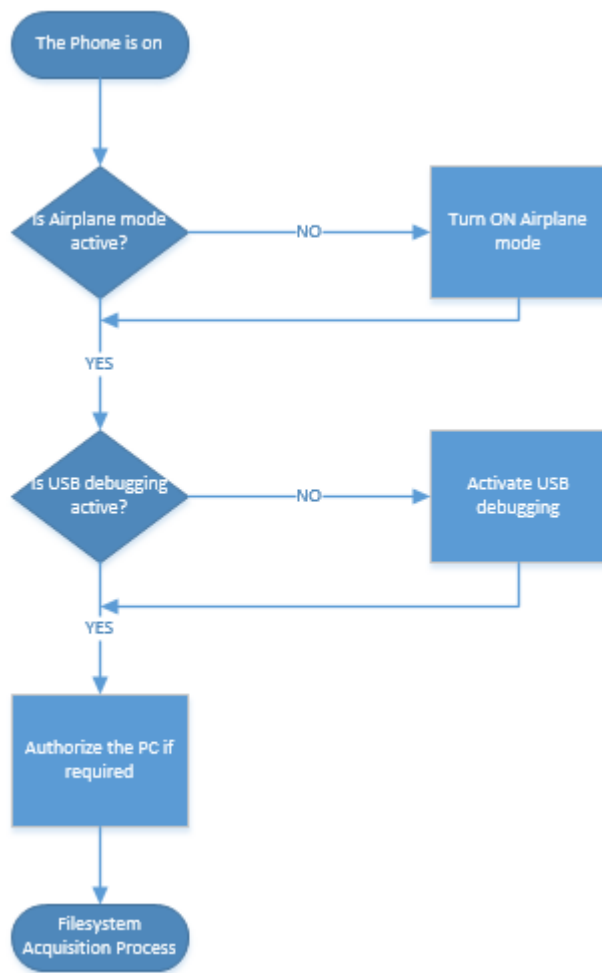
Figure 4.1: Filesystem Acquisition Process

4.2.2.4   Examination and Analysis

This section involves examining and analyzing the artifacts recovered by the paraben E3: DS

forensic tool. Table 4.2.3 below shows the location of the folder housing the most relevant artifacts that

were found in the internal memory of the Android device.

Artifact	Location	Content
		Chats
		Account info
		Friends
	/data/com.snapch at.Android/	Sent Snaps
Snapchat		Received Snaps

		Recovered Data

Table 4.2.3: Location of Artifacts Retrieved

4.3   Results

Paraben E3: DS is a forensic tool that supports app data parsing in many devices. With its inbuilt SQLite viewer, its keyword search option can be used. It also enables the user to analyze artifacts on different applications individually. In this case, a report on the Snapchat application was solely generated. The report showed some of the activities performed earlier which consist of chat messages, user account information, friends list, sent snaps, received snaps and recovered data. The account information consists of important information like account name, date of birth, the total number of snaps sent, the total number of snaps received, snap score and the username. The friend list has the names of the friends added together with a timestamp. Artifacts of the sent and received snaps were also found with timestamps and status (sent and opened or received and viewed) exclusive of the snap type. More so, chat artifacts with friends were also retrieved with timestamps and type (images or videos). A personal story that was deleted was also recovered with a timestamp. Some of the disadvantages of Paraben E3: DS are:

- Not all artifacts were present

- A visual preview of the snaps was not supported

Table 4.3.1 below shows the count of the recovered Snapchat artifacts.

User		1
Friend		5
Sent	Picture	0
	Video	0
	Message	3
Received	Picture	0
	Video	0
	Message	6
Story	Picture	2
	Video	1
	Message	1

	Audio	0
Placed Calls	Video	0
	Audio	0
Received Calls	Video	0

Table 4.3.1: Count of Artifacts Retrieved

Figure 4.2, 4.3, 4.4 and 4.5 show the received snap, sent snap, chat and the deleted story artifacts

respectively.

Received Snaps	
Time	3/17/2018 5:40:48 PM
Sender	teamsnapchat
Status	Received and viewed
Display Time (sec)	0
Date Viewed	3/17/2018 5:40:48 PM
Is Paid To Replay	No

Figure 4.2: Received Snap Artifact

Sent Snaps	
Time	4/2/2018 3:55:04 PM
Recipient	nottynerd
Status	Sent and opened

Figure 4.3: Sent Snap Artifact

Time	4/2/2018 3:58:52 PM
Sender	nottynerd
Recipients	wimmlab3
Type	Message
Text	What package?
Was seen	No

Figure 4.4: Chat Artifact

Recovered Personal Story	
Time	4/2/2018 4:01:00 PM
Display Time (sec)	3
Type	Image
Expiration Date	4/3/2018 4:01:00 PM

Figure 4.5: Deleted Story Artifact

4.4   Conclusion

In this study, the Snapchat application was analyzed to see if activities performed on it can be found

in the internal memory of an Android smartphone using the Paraben E3: DS forensic tool. The forensic tool

was not successful at retrieving the sent/received pictures and videos through chatting with friends. It was also not successful at retrieving the audio and video call exchange between the user and the friends. Fortunately, the Paraben tool was able to retrieve chat messages between the user and the friends with timestamps. Most importantly a deleted snap story was also recovered with timestamp but without a visual display of the image snap story that was deleted. Although some of the artifacts were not found, those that were recovered could be of great help during a crime investigation. For future works, this analysis could be tried on an iOS device as the Snapchat application on an Android device and iOS device operates differently. At the end of this research work, it can be concluded that some of the activities performed on the Snapchat application are stored in the internal memory of an Android smartphone.

4.5   Future Work

In this study, we examined the Snapchat application on an Android smartphone to determine if activities performed could be retrieved using the Paraben E3: DS forensic tool. Limitations of our prior studies focused solely on mobile devices. To address this limitation, in the next study we employ cloud-based predictive analytics and data science to crime data. Thus far, our studies have focused only on small data sets and could be expanded to Big Forensic Data. Previous studies such as (Al Sadi, Wimmer, Chen, and Wang (2017); Sachdev, Chen, and Rebman (2018)), did not address the power of cloud computing for analyzing Big Forensic Data. To address this, in the next study we employ the use of big data forensic tools on Microsoft Azure and IBM Cloud to analyze crime data.

5   STUDY 3: CLOUD-BASED MACHINE LEARNING OF CRIME DATA

5.1   Introduction

Cloud computing is a dominant technology that is used in executing comprehensive and complex computing. The rate at which data is increasing can be attributed to the existence of social networking applications, multimedia, and the Internet of Things (IoT). This growth has motivated researchers to design proper cloud computing platforms for analyzing this data. Big data basically denotes the large volume of data that are so challenging to store, manage and process through customary database technologies. This has led to companies adopting cloud computing due to the need to store, process and analyze these datasets. More so, characteristics such as flexibility and availability have been a major advantage in acquiring cloud computing by an organization as it allows them to focus more on the core business. The cloud service models are divided into three categories which are Platform as a Service (PaaS), Software as a Service (SaaS) and Infrastructure as a Service (IaaS). Our major concern in this study is the SaaS which refers to cloud providers hosting applications that can be accessed by users over the internet (Hashem et al., 2015). In this study, we adopted the Microsoft Azure and IBM machine learning service due to their add-and-drop feature. The add-and-drop feature makes it easy to use as it does not require deep knowledge in machine learning and programming to execute machine learning jobs on both cloud platforms making it an ideal platform for both business and forensic investigations.

Microsoft Azure is a cloud platform that houses many cloud services that organizations use in solving problems to make good business decisions. The services on the Azure cloud platform may be used alone or combined depending on the problems at hand (Azure). Some of the prevalent services provided by Microsoft Azure includes the Internet of Things (IoT), Artificial Intelligence (AI), the blockchain, mobile, e-commerce, machine learning, big data analytics, digital media and so on (Azure, 2019b). Big data problems involve the inability to store, process and analyze data and have led to the introduction of cloud computing. A major service offered by Microsoft Azure cloud is the Microsoft Azure Machine Learning Studio which is a software as a service. The Azure Machine Learning Studio is a collaborative tool with an add-and-drop feature that can be used to build, test and deploy predictive analytics solutions on datasets provided by the end user (Azure, 2019a). The interactive workspace provided by the Azure machine

learning studio aids the development of a predictive analysis model. The development of a predictive model is an iterative process which involves transforming and analyzing data from one or more data sources through several data manipulation and statistical functions so as to produce a set of results (Azure, 2019c). With the interactive and visual workspace provided by Microsoft Azure Studio, users can drag and drop datasets and the different analysis modules needed to build, test and train a predictive analysis model. These modules can be connected to form an experiment which at the end is run on the Azure machine learning studio workspace. The Azure machine learning has a plethora of modules and machine learning algorithms that can aid data input, output, visualization, and preparation. The Azure machine learning studio is known for its four well-known categories of algorithms which are classification, regression, clustering and anomaly detection. The building, testing and running of a predictive model in the Azure machine learning studio is time efficient as it takes less time to perform and predictive analytic experiment.

IBM Cloud is a cloud computing platform by IBM that offers a platform as a service (PaaS), software as a service (SaaS) and infrastructure as a service (IaaS). Organizations can access and deploy virtualized IT resources with infrastructure as service on the IBM cloud (IBM, 2019a). The IaaS provided by IBM cloud is not limited to networking, compute power and storage of data over the internet. The PaaS in IBM cloud aids developers in creating, managing, running and deploying lots of applications in the public cloud. Also, various types of programming languages such as Python, Java, PHP, and Node.js are supported across the IBM cloud (IBM, 2019d). IBM cloud offers many services like compute, network, storage, management, security, analytics, Artificial Intelligence (AI), Internet of Things (IoT), the blockchain, mobile and so on (IBM, 2019b). More so, the IBM cloud platform provides access to other IBM tools and services including the IBM Watson. Watson is a top performance question-answering supercomputer made by IBM that functions through the combination of analytical software and Artificial Intelligence (AI). The supercomputer was named after the IBM founder, Thomas J. Watson. The IBM Watson Studio delivers tools for data scientists and application developers to work with data in order to build and train models at large scale (IBM, 2019c). The IBM Watson SPSS Modeler offers an add-and-drop feature like the Microsoft Azure Studio. This feature makes it easy for users who are new to machine learning to easily connect the several modules needed to train and deploy their predictive model. The IBM Watson Studio suites lots of machine learning classifiers like a decision tree, Linear regression, logistic regression, neural network,

support vector machine, random forest, random trees and so on.

5.2 Methodology

This section of the study was to perform predictive analysis using Microsoft Azure and IBM cloud platforms. The predictive analysis was performed on both Azure Machine learning Studio and IBM Watson Studio using three different datasets from the UCI machine learning repository and Kaggle. The datasets are default credit card payment, appointment no-show and the Titanic dataset. More so, three different machine learning algorithms were used on both cloud platforms to see how well they perform in order to select a cloud tool to run the main analysis. After performing several experiments on each platform using these algorithms: Logistic Regression (LR), Artificial Neural Network (ANN) and Decision Tree (DT), the results of the experiments can be seen in the tables and figures below. The results achieved on both platforms are kind of similar as well as their pricing which is $0.5 USD per 1000 predictions. They both offer a free trial for users to test the tools, but limited experiments can be performed on the free tier. Based on the ease of use, real-time predictions and flexibility, the Azure Machine Learning Studio was selected to build and train the predictive analysis model for this study.

The tables 5.3.1, 5.3.2 and 5.3.3 below shows the results of the experiments performed on Microsoft Azure and IBM Cloud using the appointment no-show, default credit card payment and the titanic datasets respectively.

	Logistic Regression		Decision Tree		Artificial Neural Network	
	Azure	IBM	Azure	IBM	Azure	IBM
Accuracy	79.30%	79.49%	79.50%	80.08%	79.30%	79.59%
AUC	0.665	0.655	0.737	0.603	0.712	0.68

Table 5.2.1: Model Evaluation for Appointment No-show Dataset

	Logistic Regression		Decision Tree		Artificial Neural Network	
	Azure	IBM	Azure	IBM	Azure	IBM
Accuracy	81.20%	80.20%	81.00%	80.12%	82.20%	80.80%
AUC	0.723	0.722	0.709	0.715	0.764	0.754

Table 5.2.2: Model Evaluation for Default Credit Card Payment Dataset

	Logistic Regression		Decision Tree		Artificial Neural Network	
	Azure	IBM	Azure	IBM	Azure	IBM

Accuracy	77.90%	77.42%	79.80%	78.01%	78.88%	78.59%
AUC	0.785	0.739	0.798	0.71	0.798	0.748

Table 5.2.3: Model Evaluation for Titanic Dataset

Figures 5.1 to 5.8 below shows predictive model workflow and the respective results that were achieved using the Artificial Neural Network (ANN), Logistic Regression (LR) and Decision Tree (DT) classifiers on the appointment no-show dataset using the Microsoft Azure Machine Learning Studio and the IBM Watson SPSS Modeler.
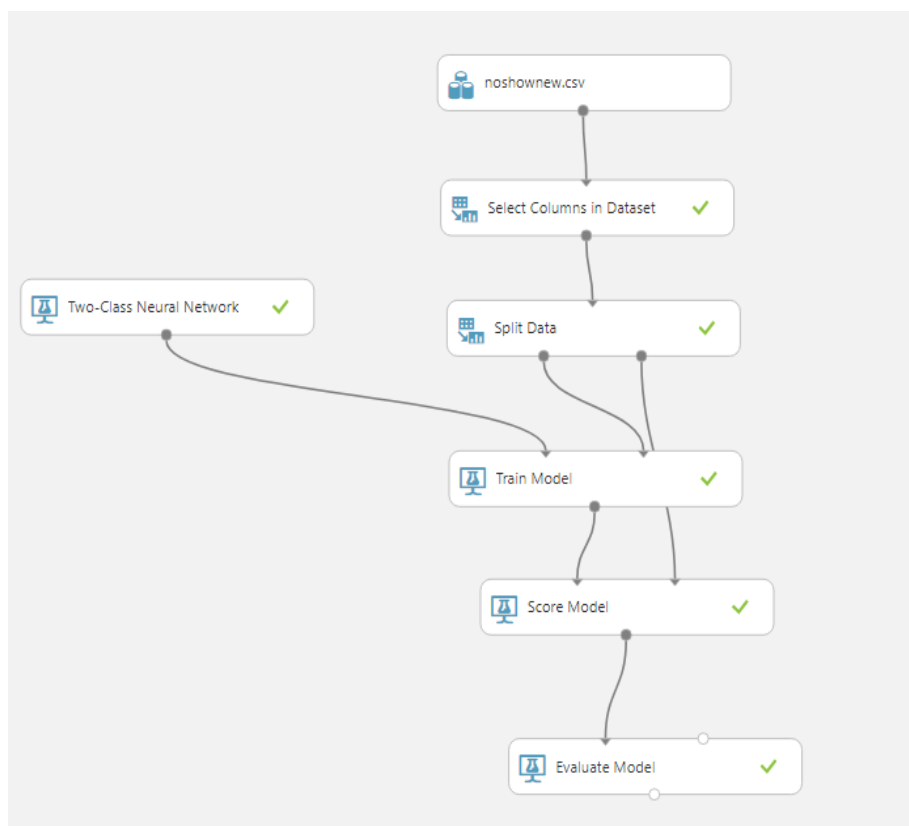
Figure 5.1: Azure ANN Model for Appointment No-show

ROC PRECISION/RECALL LIFT

True Positive	False Negative	Accuracy	Precision	Threshold	AUC
250	6502	0.793	0.419	0.5	0.712
False Positive	True Negative	Recall	F1 Score		
347	26059	0.037	0.068		
Positive Label	Negative Label				
Yes	No				
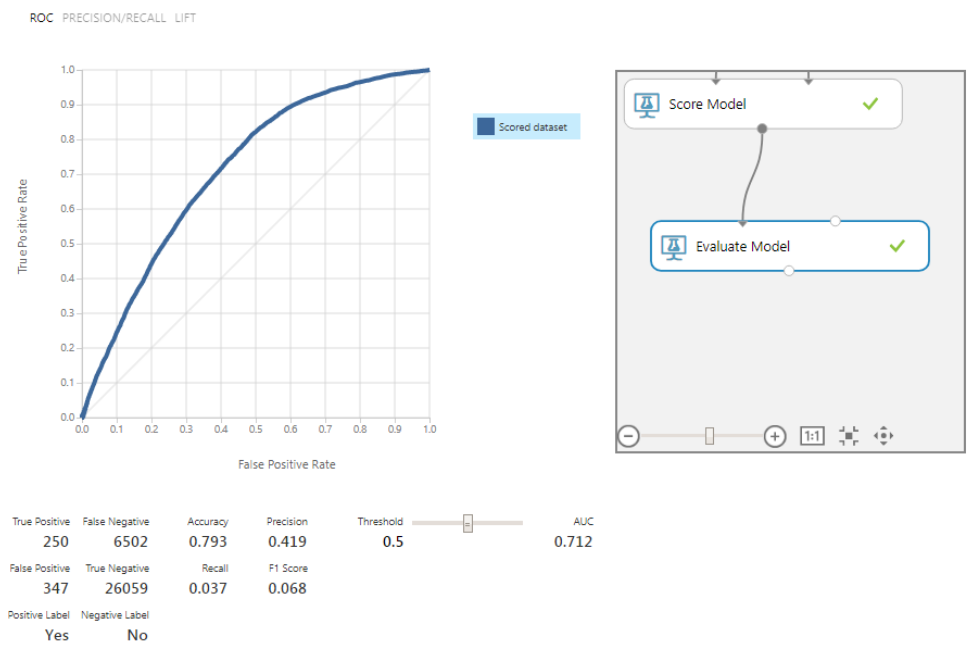
Figure 5.2: Azure ANN Result for Appointment No-show

ROC PRECISION/RECALL LIFT

True Positive	False Negative	Accuracy	Precision	Threshold	AUC
93	6659	0.793	0.330	0.5	0.665
False Positive	True Negative	Recall	F1 Score		
189	26217	0.014	0.026		
Positive Label	Negative Label				
Yes	No				

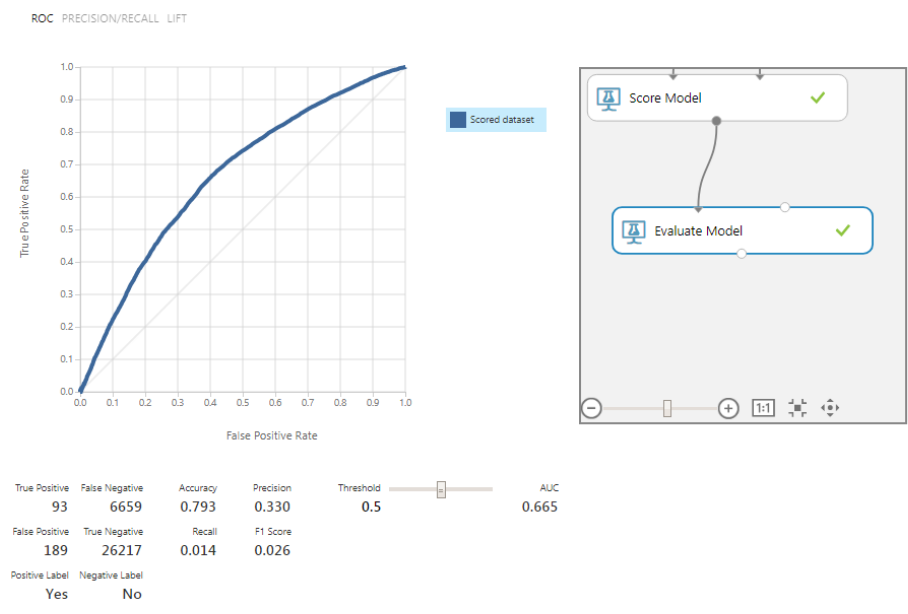*Figure 5.3: Azure Logistic Regression Result for Appointment No-show*

Patient Appointment No Show › Evaluate Model › Evaluation results

ROC PRECISION/RECALL LIFT

True Positive	False Negative	Accuracy	Precision	Threshold	AUC
860	5892	0.795	0.484	0.5	0.737

False Positive	True Negative	Recall	F1 Score		
917	25489	0.127	0.202		

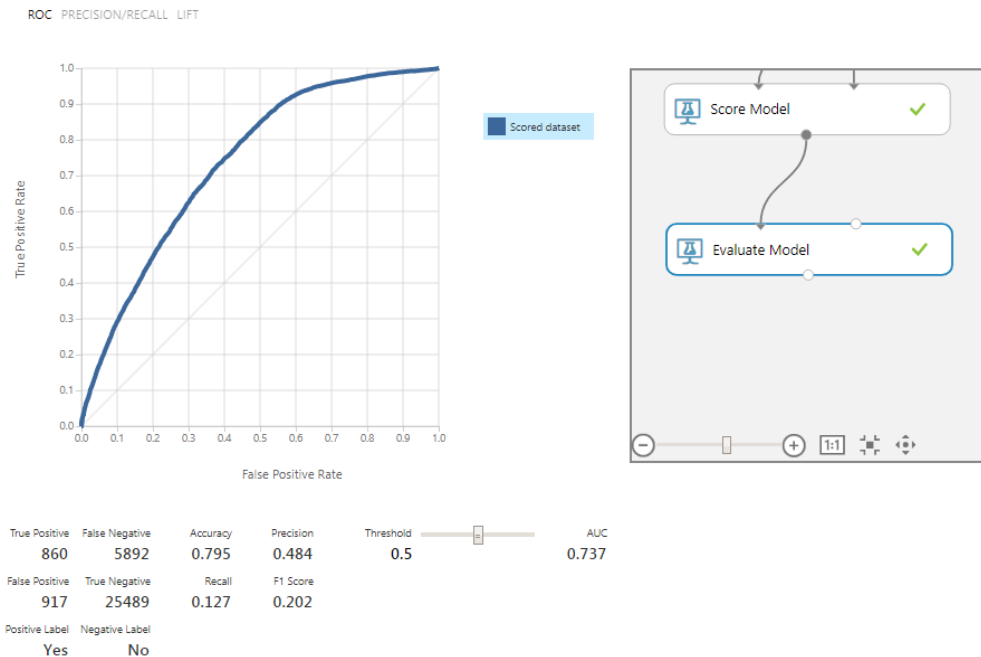| Positive Label	Negative Label				
Yes	No				

Figure 5.4: Azure Decision Tree Result for Appointment No-show

Figure 5.5: IBM ANN Model for Appointment No-show

Results for output field No-show
Individual Models
Comparing $C-No-show with No-show

'Partition'	1_Training		2_Testing	
Correct	61,844	80.21%	26,770	80.08%
Wrong	15,254	19.79%	6,659	19.92%
Total	77,098		33,429	

Performance Evaluation

'Partition' = 1_Training

No	0.007
Yes	1.124

'Partition' = 2_Testing

No	0.009
Yes	1.153

Evaluation Metrics

'Partition'	1_Training		2_Testing	
Model	AUC	Gini	AUC	Gini
$C-No-show	0.595	0.19	0.603	0.205

Figure 5.6: IBM Decision Tree Result for Appointment No-show
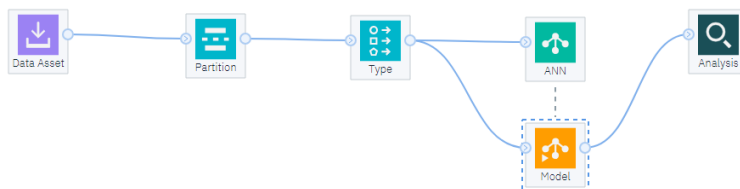
```
Results for output field No-show
Individual Models
Comparing $L-No-show with No-show
'Partition'              1_Training                2_Testing
Correct                  61,433    79.68%          26,573   79.49%
Wrong                    15,665    20.32%           6,856   20.51%
Total                    77,098                    33,429
Performance Evaluation
'Partition' = 1_Training
No                                   0.001
Yes                                  0.444
'Partition' = 2_Testing
No                                   0.001
Yes                                  0.619
Evaluation Metrics
'Partition'              1_Training                2_Testing
Model                       AUC    Gini              AUC   Gini
$L-No-show                0.653   0.307            0.655   0.31
```

Figure 5.7: IBM Logistic Regression Result for Appointment No-show

```
Results for output field No-show
Individual Models
Comparing $N-No-show with No-show
'Partition'              1_Training                2_Testing
Correct                  61,585    79.88%          26,607   79.59%
Wrong                    15,513    20.12%           6,822   20.41%
Total                    77,098                    33,429
Performance Evaluation
'Partition' = 1_Training
No                                   0.0
Yes                                  0.859
'Partition' = 2_Testing
No                                   0.0
Yes                                  0.52
Evaluation Metrics
'Partition'              1_Training                2_Testing
Model                       AUC    Gini              AUC   Gini
$N-No-show                0.678   0.356             0.68   0.361
```

Figure 5.8: IBM Neural Network Result for Appointment No-show

5.2.1 Data Preparation

This section involves merging two datasets which will later be used in the predictive analysis. The first dataset is the Chicago crime dataset which is gotten from the data.gov repository. It has 21 attributes and over a million records. The second dataset is the Chicago Department of Public Health social economic dataset from data.gov repository. It has 9 attributes and 77 records. With these two datasets, I was able to merge them through a similar attribute which is the community area number using a python library known as Pandas. The description of the two datasets can be seen in Tables 5.3.4 and 5.3.5 below.

Attribute	Data Type
Case Number	String

Date	Date time
Block	String
IUCR	Integer
Primary Type	String
Description	String
Location Description	String
Arrest	Binary
Domestic	Binary
Beat	Integer
District	Integer
Ward	Integer
Community Area Number	Integer
FBI code	String
X Coordinate	Integer
Y Coordinate	Integer
Year	Date
Updated On	Date time
Latitude	Float
Longitude	Float
Community Area Name	String

Table 5.2.4: Chicago Crime Dataset

Attribute	Data Type
Percent of housing crowded (POHC)	float
Percent of households below poverty (PHBL)	float
Percent aged 16+ unemployed (PAU16)	float

|	
Percent aged 25+ without a high school diploma (PA25WHD)	float
Percent aged under 18 and over 16 (PAUO)	float
Per capita income (PCI)	Integer
Hardship index (HI)	Integer
Community Area Number	Integer
Community Area Name	String

Table 5.2.5: Social Economic Dataset

5.2.2   Choosing a Model

As this study is aimed at discovering a relationship between the dependent attribute and the independent attributes in the merged dataset, linear regression can be applied. Linear regression is an uncomplicated and frequently used statistical method that can be used in analyzing the relationships and association between a target variable and other variables in order to make accurate predictions.

5.2.3   Environment Setup

The training of the merged dataset is done using the Azure machine learning studio. The dataset is loaded from a local source into the Azure machine learning environment as seen in the Table 5.3.6 below. Afterward, the machine learning experiment can then be performed.
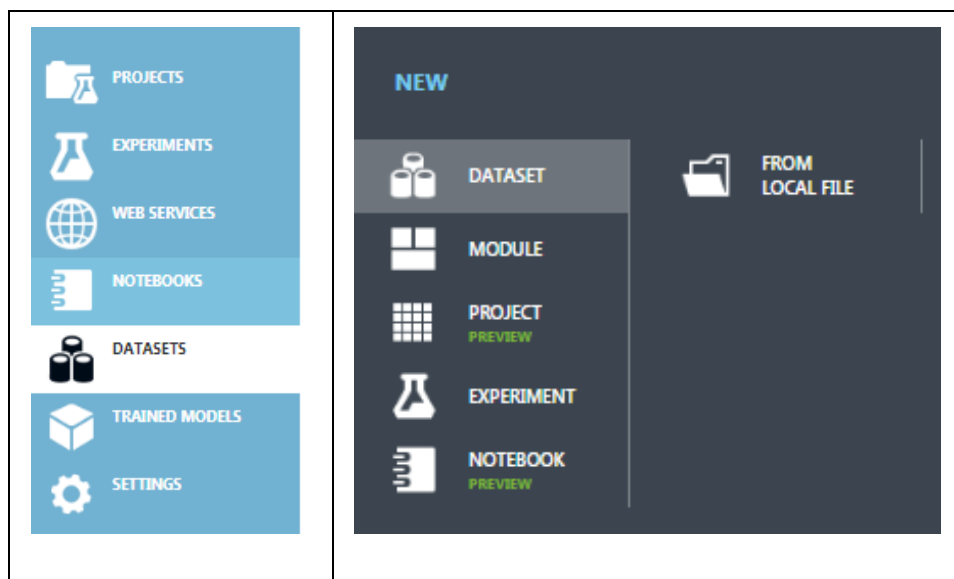
Table 5.2.6: The Dataset Section of the Azure ML Studio

5.2.4   Experiment

After loading the dataset, the experiment can then be performed by following some specific steps as seen in the flowchart in Figure 5.9 below.
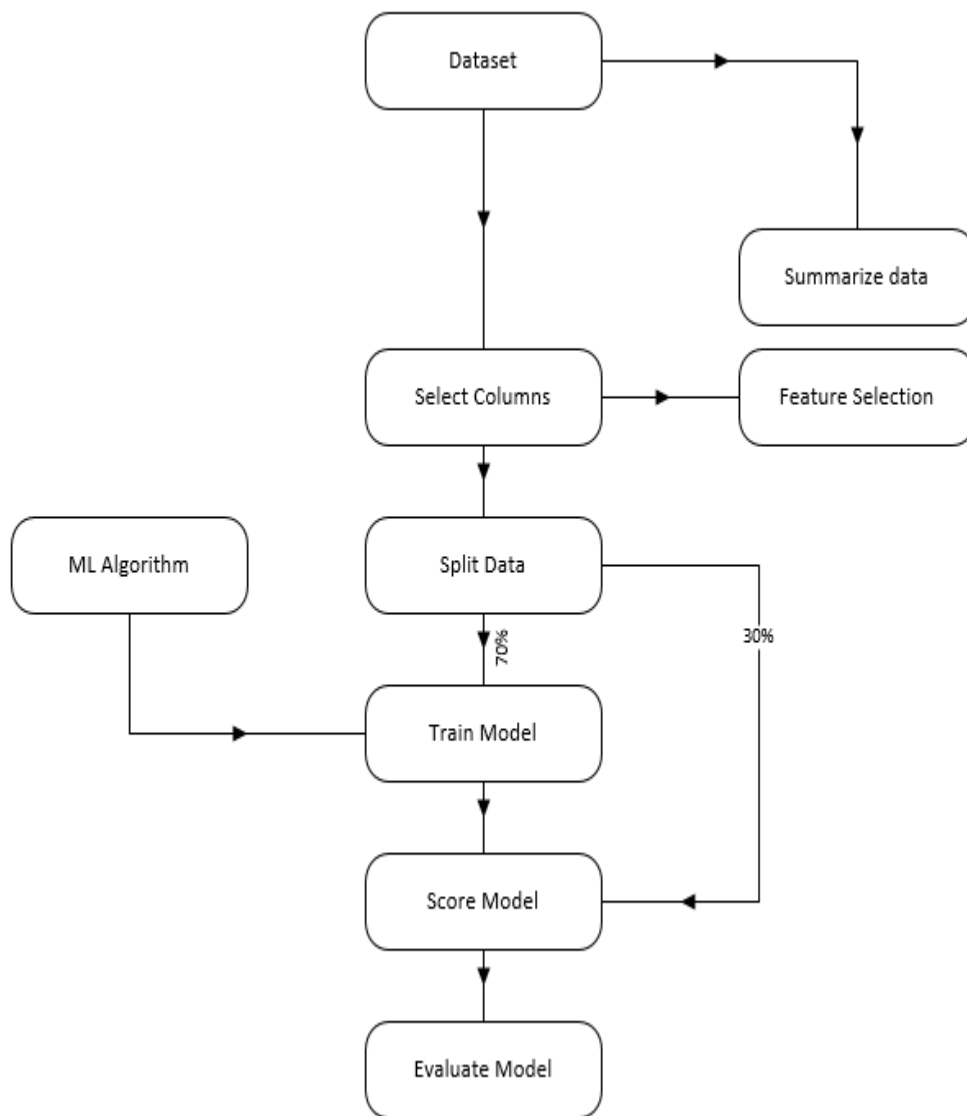
Figure 5.9: Machine Learning Flowchart in Azure Studio

With the drag and drop feature in the Azure machine learning studio, the different modules in the flowchart above can be dragged into the graphical interface. The modules in the flowchart are explained below:

5.2.4.1   Dataset

The dataset in the flowchart chart above is the dataset that was loaded into the environment for data analysis. This is then connected to the "summarize data" module.

5.2.4.2   Summarize Data

This module gives the summary statistics such as mean, minimum value, maximum value, count, unique value count, missing value count, range, median and mode of each column in the dataset. Doing this is important to know if there are missing values as well as checking if one or more input variable dominates others as these may affect the accuracy of the trained model. Since the dataset to be used in our analysis has no missing values and the inputs do not need normalization, the module containing the dataset is then connected to the select columns module.

5.2.4.3   Select Data Columns

This is where the columns required for our analysis is selected. Originally, there are 10 columns but 8 will be selected as the community number and community names are not predictive columns. The columns selected for our analysis are percent of housing crowded (POHC), percent household below poverty (PHBP), percent aged 16+ unemployed (PA16U), percent aged 25+ without high school diploma (PA25WHD), percent aged under 18 or over 64 (PAUO), per capita income (PCI), hardship index (HI) and crime rate. A connection can now be made from the select columns module to the split data module.

5.2.4.4   Split Data

The next step is splitting the dataset into two portions which are training and the testing portions. The splitting is done in a randomized manner to allow good accuracy. In this study, we will be splitting the dataset in the ratio 70:30. That is, 70 percent of the dataset will be used for training the model while the remaining 30 percent will be used to test the model. As seen in the flowchart above, 70 percent of the dataset is connected directly to the train model module. The test dataset will be used later in our analysis.

5.2.4.5   Machine Learning Algorithm

Linear regression is the machine learning algorithm used in this study as we are looking to see if there is a relationship between the independent variables and the dependent variable. The ordinary least squares (OLS) method was selected as the solution method to measure error and fit the regression line. OLS is a type of linear least squares method used for estimating unknown parameters by minimizing the squared error in a linear regression model (Azure). Afterward, the linear regression module is dragged to the

graphical panel and then connected directly to the train model module (Azure).

5.2.4.6   Train Model

The column to be predicted which is also known as the target variable must be selected in the train model module. In this study, the target variable we are trying to predict is the overall crime rate in each community in Chicago.

5.2.4.7   Score Model

This module is where the trained model is scored and tested with the remaining 30 percent of the dataset that was set aside. It can be seen from the flowchart that the testing dataset is connected directly to the score model module.

5.2.4.8   Evaluate Model

This module is connected directly to the score module as it displays the performance of the model.

5.2.4.9   Permutation Feature Importance

The permutation feature importance model has two inputs and is connected to the output of the trained model and the test data output of the split data module. This helps in discovering how important the dependent features are to the target variable. The metric used for measuring their importance is the coefficient of determination which is also known as the R squared value. Figure 5.10 below shows the Linear Regression (LR) flow diagram in Microsoft Azure Machine Learning Studio.
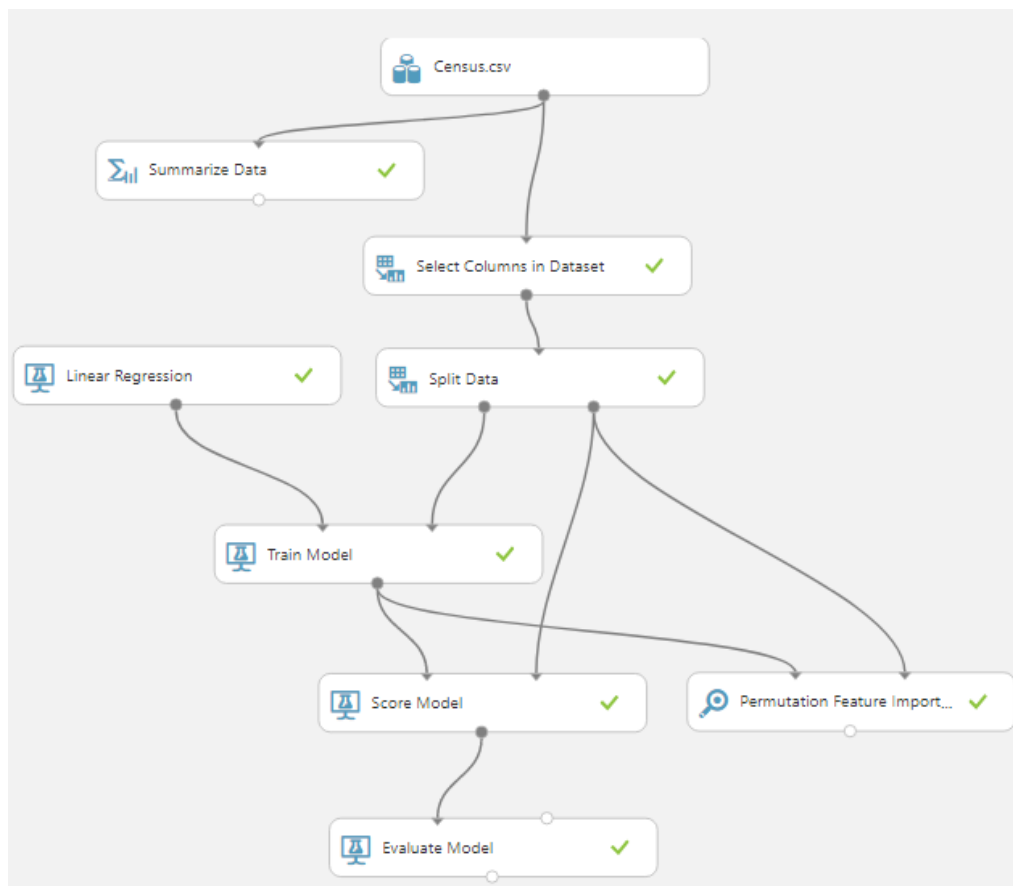
Figure 5.10: Linear Regression Flow Diagram

5.3   Results

This phase of the study encompasses the performance of our trained model. Table 5.4.1 below shows the result obtained after running the analysis on the Azure machine learning studio.

Metrics	
Mean Absolute Error	13125.55801
Coefficient of Determination	0.22322

Table 5.3.1: Evaluation Metrics

The Mean Absolute Error (MAE) measures the closeness between the predicted values and the actual values. This is achieved by calculating the prediction error for each of our test data. Afterward, they are converted to their absolute value. Finally, the mean which is the average sum of all the absolute errors in our test data is then calculated. The result of our analysis signifies on the average, the model predictions are off by approximately 13125 which approximately equals 11%. A low value of mean absolute error signifies a good model.

The Coefficient of Determination also known as the R squared (R^2) value denotes the predictive power of the model and it's usually between 0 and 1. It is used in explaining how variability can be caused by the relationship between attributes. When a value of Zero is achieved, it simply means the model is random while 1 indicates a perfect fit. Nevertheless, we should be careful when interpreting the R^2 values as lowers values can entirely be normal and high values can be otherwise. The value of R-squared in this model is 0.223 which means that around 22% of the variation in crime rate is explained by the model.

5.4   Conclusion

The goal of this study is to perform machine learning in the cloud. Microsoft Azure Machine Learning Studio and IBM Watson SPSS Modeler were initially compared to see how well they perform and their ease of use. Different machine learning techniques were used on several datasets across the two cloud services to see how well they perform. Although their performances and usage were quite similar based on the friendly graphical interface and the "add and drop" feature. The "add and drop" feature makes it easy for users with little or no experience in machine learning to perform data analysis. Also, performing data analysis of both cloud services does not require writing codes compared to other cloud providers. In the end, Microsoft Azure Machine Learning Studio was selected for its speed at performing data analysis compared with the IBM Watson SPSS Modeler which takes a longer time at performing the same analysis. Nevertheless, the aim of this research is seeing how social economic indexes in Chicago affect the crime rates in the cities of Chicago. The result of running a linear regression algorithm on Microsoft Azure Machine Learning Studio shows that hardship index, percent aged under 18 or over 64 and percent aged 25+ without high school diploma are the most important factors at predicting the crime rate in each of the communities in Chicago. More so, knowing these social-economic indexes identified during our data analysis could aid the government and police department in making the best decisions at tackling the various crimes occurring in the cities of Chicago.

6   CONCLUSIONS

Due to the rate at which criminal activities are being carried out using smartphones, mobile forensics have been able to help in tackling this issue. The first two studies in this research work have shown how forensically accepted tools and techniques can be used to analyze data on a smartphone. The results of data analysis performed in study one and two could assist forensic investigators during the forensic investigation. From study one, more artifacts were retrieved when a commercial forensic tool was used compared to when the open source tool was used. Activities performed using WhatsApp, Twitter, Google Drive, and Facebook applications were all retrieved with timestamps using Paraben E3: DS which is a commercial tool. Moreover, when a commercial forensic tool was used in study two, not all the predefined activities performed on the Snapchat application were retrieved. Artifacts like sent/received pictures and videos between users were not retrieved. Fortunately, chat messages and a deleted snap story was retrieved with timestamps. Unfortunately, the Paraben E3: DS was not successful at retrieving the audio and video call exchange between the user and friends.

Cloud computing is a dominant technology that is used in executing comprehensive and complex computing. Due to the urgent need for data to be stored, processed and analyzed, big data problems can be solved using cloud computing services. The "add and drop" feature on both cloud tools makes it easy for users with little or no experience in machine learning to perform data analysis. The result of running a linear regression algorithm on the Microsoft Azure Machine Learning Studio shows that hardship index, percent aged under 18 or over 64 and percent aged 25+ without high school diploma are the most important factors at predicting the crime rate in each of the communities in Chicago. More so, knowing these social-economic indexes identified during our data analysis could aid the government and police department in making the best decisions at tackling the various crimes occurring in the communities in Chicago.

REFERENCES

Abdelaziz, A., Elhoseny, M., Salama, A. S., & Riad, A. (2018). A machine learning model for improving healthcare services on cloud computing environment. *Measurement, 119*, 117-128.

Al Sadi, M. B., Wimmer, H., Chen, L., & Wang, K. (2017). *Improving the Efficiency of Big Forensic Data Analysis Using NoSQL.* Paper presented at the Proceedings of the 10th EAI International Conference on Mobile Multimedia Communications.

Anglano, C. (2014). Forensic analysis of WhatsApp Messenger on Android smartphones. *Digital Investigation, 11*(3), 201-213.

Autopsy Forensic Tool. from https://www.sleuthkit.org/autopsy/

Awan, F. A. (2015). *Forensic examination of social networking applications on smartphones.* Paper presented at the Information Assurance and Cyber Security (CIACS), 2015 Conference on.

Ayers, R., Brothers, S., & Jansen, W. (2013). Guidelines on mobile device forensics (draft). *NIST Special Publication, 800*, 101.

Aziz, N. A., Mokhti, F., & Nozri, M. N. M. (2015). *Mobile Device Forensics: Extracting and Analysing Data from an Android-based Smartphone.* Paper presented at the Cyber Security, Cyber Warfare, and Digital Forensic (CyberSec), 2015 Fourth International Conference on.

Azure, M. Linear Regression. from https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/linear-regression

Azure, M. What is Azure?   Retrieved 01/23/2019, 2019, from https://azure.microsoft.com/en-us/overview/what-is-azure/

Azure, M. (2019a). Azure Machine Learning Studio. from https://docs.microsoft.com/en-us/azure/machine-learning/studio/

Azure, M. (2019b). Microsoft Azure Solutions. from https://azure.microsoft.com/en-us/solutions/

Azure, M. (2019c). What is ML Studio from https://docs.microsoft.com/en-us/azure/machine-learning/studio/what-is-ml-studio

Bewick, V., Cheek, L., & Ball, J. (2005). Statistics review 14: Logistic regression. *Critical care, 9*(1), 112.

Burke Winkelman, S., Oomen-Early, J., Walker, A. D., Chu, L., & Yick-Flanagan, A. (2015). Exploring Cyber Harassment among Women Who Use Social Media. *Universal Journal of Public Health, 3*(5), 194.

Chen, C. P., & Zhang, C.-Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences, 275*, 314-347.

Chu, H.-C., Lo, C.-H., & Chao, H.-C. (2013). The disclosure of an Android smartphone's digital footprint respecting the Instant Messaging utilizing Skype and MSN. *Electronic Commerce Research, 13*(3), 399-410.

Chu, H.-C., Yang, S.-W., Wang, S.-J., & Park, J. H. (2012). The partial digital evidence disclosure in respect to the instant messaging embedded in viber application regarding an android smart phone *Information technology convergence, secure and trust computing, and data management* (pp. 171-178): Springer.

Cusumano, M. (2010). Cloud computing and SaaS as new computing platforms. *Communications of the ACM, 53*(4), 27-29.

Dogan, S., & Akbal, E. (2017a). *Analysis of mobile phones in digital forensics.* Paper presented at the Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2017 40th International Convention on.

Dogan, S., & Akbal, E. (2017b). *Analysis of mobile phones in digital forensics.* Paper presented at the 2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO).

Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: a methodology review. *Journal of biomedical informatics, 35*(5-6), 352-359.

Farnden, J., Martini, B., & Choo, K.-K. R. (2015). Privacy risks in mobile dating apps. *arXiv preprint arXiv:1505.02906*.

Gai, K., & Li, S. (2012). *Towards cloud computing: a literature review on cloud computing and its development trends.* Paper presented at the Multimedia Information Networking and Security (MINES), 2012 Fourth International Conference on.

Gao, F., & Zhang, Y. (2013). *Analysis of WeChat on iPhone.* Paper presented at the 2nd International Symposium on Computer, Communication, Control, and Automation (3CA).

Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Khan, S. U. (2015). The rise of "big data" on cloud computing: Review and open research issues. *Information systems, 47*, 98-115.

Huang, C.-L., Chen, M.-C., & Wang, C.-J. (2007). Credit scoring with a data mining approach based on support vector machines. *Expert systems with applications, 33*(4), 847-856.

Husain, M. I., & Sridhar, R. (2009). *iForensics: forensic analysis of instant messaging on smart phones.* Paper presented at the International Conference on Digital Forensics and Cyber Crime.

IBM. (2019a). IBM Cloud. from https://www.ibm.com/cloud/

IBM. (2019b). IBM Cloud prodeucts. from https://www.ibm.com/cloud/products

IBM. (2019c). IBM Watson Studio. from https://www.ibm.com/cloud/watson-studio

IBM. (2019d). Watson SDKs. from https://console.bluemix.net/docs/services/watson/getting-started-sdks.html#using-sdks

Lee, C., & Chung, M. (2015). Digital Forensic Analysis on Window8 Style UI Instant Messenger Applications *Computer Science and its Applications* (pp. 1037-1042): Springer.

Lessard, J., & Kessler, G. (2010). Android Forensics: Simplifying Cell Phone Examinations.

Levinson, A., Stackpole, B., & Johnson, D. (2011). *Third party application forensics on apple mobile devices.* Paper presented at the System Sciences (HICSS), 2011 44th Hawaii International Conference on.

Li, A., Yang, X., Kandula, S., & Zhang, M. (2010). *CloudCmp: comparing public cloud providers.* Paper presented at the Proceedings of the 10th ACM SIGCOMM conference on Internet measurement.

Lukito, N. Y. P., Yulianto, F. A., & Jadied, E. (2016). *Comparison of data acquisition technique using logical extraction method on Unrooted Android Device.* Paper presented at the

Information and Communication Technology (ICoICT), 2016 4th International Conference on.

Mahajan, A., Dahiya, M., & Sanghvi, H. (2013). Forensic analysis of instant messenger applications on android devices. *arXiv preprint arXiv:1304.4915*.

Mell, P., & Grance, T. (2011). The NIST definition of cloud computing.

Norouzizadeh Dezfouli, F., Dehghantanha, A., Eterovic-Soric, B., & Choo, K.-K. R. (2016). Investigating Social Networking applications on smartphones detecting Facebook, Twitter, LinkedIn and Google+ artefacts on Android and iOS platforms. *Australian journal of forensic sciences, 48*(4), 469-488.

Paraben Corporation. from https://shop.paraben.com/index.php?id_product=133&controller=product

Qiu, J., Wu, Q., Ding, G., Xu, Y., & Feng, S. (2016). A survey of machine learning for big data processing. *EURASIP Journal on Advances in Signal Processing, 2016*(1), 67.

Roy, N. R., Khanna, A. K., & Aneja, L. (2016). *Android phone forensic: Tools and techniques.* Paper presented at the Computing, Communication and Automation (ICCCA), 2016 International Conference on.

Sachdev, H., Chen, L., & Rebman, C. (2018). A New Framework for Securing, Extracting and Analyzing Big Forensic Data. *Journal of Digital Forensics, Security and Law, 13*(2), 6.

Thakur, N. S. (2013). Forensic analysis of WhatsApp on Android smartphones.

Thomas, P., Owen, P., & McPhee, D. (2010). *An analysis of the digital forensic examination of mobile phones.* Paper presented at the Next Generation Mobile Applications, Services and Technologies (NGMAST), 2010 Fourth International Conference on.

Walker, L. (2017). How to Use Snapchat: Share Vanishing Photos with Snap Chat. from https://www.lifewire.com/how-to-use-snapchat-2654838

Wimmer, H., Chen, L., & Narock, T. (2018). Ontologies and the Semantic Web for Digital Investigation Tool Selection. *Journal of Digital Forensics, Security and Law, 13*(3), 6.