Research on identifying Data Leak Events in an incoming payload stream of an IoT platform

Ashari, Rafidah

ABSTRACT

RESEARCH ON IDENTIFYING DATA LEAK EVENTS IN AN INCOMING PAYLOAD STREAM OF AN IOT PLATFORM

This study aims to identify data leak events from an incoming Internet of Things (IoT) data stream generated by IoT devices to the Internet Cloud of Things (CoT) platform. IoT plays an increasingly significant role in everyday life. Most private and commercial consumers interact with different kinds of smart IoT devices, collecting one or multiple data types, like presence sensors, voice-activated assistance, remote patient monitoring, and glucose level monitoring (Emami-Naeini et al., 2017). The built-in intelligence in the IoT–connected devices enables them to connect, communicate, report, and execute specific instructions automatically without human intervention. In this ecosystem, it is crucial to consider user privacy and personal data protection to avoid data breaches and to protect against malicious activities of cybercriminals. General Data Protection Regulation (GDPR) Article 4 on personal data defines types of data that need to be protected and what kind of data can identify a natural person (Vojković et al., 2020). GDPR includes another section named Privacy-by-Design (Perera et al., 2016) that mentions data protection explicitly via technology, which is the focus of this study.

This study has been executed in the context of Deutsche Telekom (DT) and, specifically, their design of IoT and CoT platforms. The main goal of this study is to propose a method for DT to identify data leaking events in their IoT pipeline so that they can deploy appropriate controls. The study investigates the feasibility of machine learning (ML)-based and rule-based (RB) detection mechanisms and evaluates these designs with two key stakeholders at DT.

This study generated the input file in the data preparation phase mimicking the IoT data payload format using a Python script. The data was trained with a supervised learning method using a tree-based machine learning classifier and then performed predictions. The result showed that the ML model could detect personally identifiable information (PII) (in this case, location data) amongst all other messages in the JSON files of the randomly generated IoT payload data. The study used the same input files from the ML method, which contains a mixture of PII and non-PII data for the RB approach. The RB method directly filtered out the location data using the regular expressions formula. The study's outcome found that machine learning and rule-based methods can identify PII. This study outcome could enhance Deutsche Telekom's 'Privacy-by-Design' incorporated in their IoT products and services. Thus, the company can proactively act on possible data breach incidents by utilising the 'data leaking alert.'

ACKNOWLEDGEMENT

I want to give special thanks to my first supervisor, Dr Olga Gadyatskaya, whose invaluable advice helped steer my thesis in the right direction. Thank you to my second supervisor, Dr Apostolis Zarras, for his feedback during the checkpoint sessions, which provided ideas for improving the thesis writing flow.

I want to express my gratitude to the company supervisor Dr Gerald Delvos, for his constant guidance. Thank you to my supportive friends and colleagues who proofread and provided valuable feedback on my thesis.

Finally, I would like to thank my husband for his patience throughout this thesis journey.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1. Introduction

1.1 Chapter Introduction

This chapter presents the motivations for the research topic. Generally speaking, the study concerns personal data leaks in the Internet of Things (IoT) ecosystem. Modern society relies on the internet for various services, such as mobile computing, web applications, internet banking, and other various applications (Sahmim et al., 2017). In this scenario, where everything connects everywhere, we face security and privacy challenges that represent the main discussion topic amongst the researchers, enforcement agencies, and governmental bodies (Le et al., 2018). IoT devices, sensors, and actuators transmit an enormous amount of end-user-related data via the Internet to reach the Cloud of Things (CoT). These data will be processed and analysed according to the needs and goals of the IoT system provider. The analysis outcome is accessible to the end users via their specific application in the cloud. From here, we can see how data flows via Internet-connected devices, including the personal and private data of a particular data owner ('Natural Person') (Cerbo et al., 2018). As private data travels to and from IoT devices, cybercriminals are able to steal this data. Worst-case scenarios include stolen identity, personal data exposure, or having the information sold on the dark web (Abomhara et al., 2015). A security or privacy breach will hurt the IoT system provider's reputation since their customers might legitimately assume that their personal information is compromised, even if this is not always the case (Kamin, 2017). There are two ways in which IoT can interact with personal data, habits, and movements (Vojković et al., 2020). First, smart devices empowered by IoT can collect more data than traditional non-connected devices. For example, in a Smart City environment where IoT devices continuously collect data related to population activities and multiple agencies have access to these data, ownership and data retention periods are difficult to establish. This has led to significant privacy and security concerns (Barnaghi et al., 2013). Second, IoT devices themselves are vulnerable to hacking and abuse by cybercriminals (Abomhara et al., 2015).

Unfortunately, the increasing usage of IoT-network-connected devices in complex information systems has created opportunities for data security breaches (Vojković et al., 2020). IoT providers need to develop solutions for detecting privacy leaks to mitigate this issue, including leaks from IoT sensors and actuators. Thus, the study will identify which part of data transmitted from IoT sensors to the CoT belongs to the personally identifiable information (PII) in the context of article 4 of the General Data Protection Regulation (GDPR)–specified PII data protection (Kaneen et al., 2020; McCallister et al., 2010). By knowing which segment of IoT data is PII in the IoT payload, IoT providers can create an alert system to monitor data leak events to proactively scan the integrity of personal data and react in case of privacy incidents. Functioning like a honeypot, this alert system can proactively alert the IT hosting or providers of any hostile activities long before the cybercriminals attack the existing infrastructure or applications (Sink, 2001). Using artificial intelligence technology, specifically machine learning (ML) and rule-based (RB) filtering of data, this study will determine which approach better recognises private and personal data and their advantages and disadvantages. In the next section, we will define the keywords used in this study. Next, we will clarify the research questions and objectives and summarise the approach, the research

methods used, and the potential challenges. Finally, we will also explain the organisation of the whole thesis and the desired outcome it intends to achieve throughout this study.

1.2 Terminology

Internet of Things (IoT) is a keyword frequently used in this study. IoT is a set of networked, connected IT devices or sensors that transmit and exchange messages over the Internet (Cloud of Things) (Sahmim et al., 2017). Machina Research forecasted about twenty-seven billion IoT connections by 2025, increasing the compound annual growth rate (CAGR) by 16% from six billion IoT connections in 2015 (Machina Research, 2016; Qu et al., 2017). From a financial perspective, McKinsey Global Institute estimated that the potential economic impact of IoT will increase to USD11.1 trillion per year in 2025 compared with USD3.9 trillion in 2015 (Manyinka et al., 2015; Vojković et al., 2020).

Figure 1 depicts an end consumers' IoT devices connected to the Internet. All these devices are transmitting and receiving messages to and from the Internet.

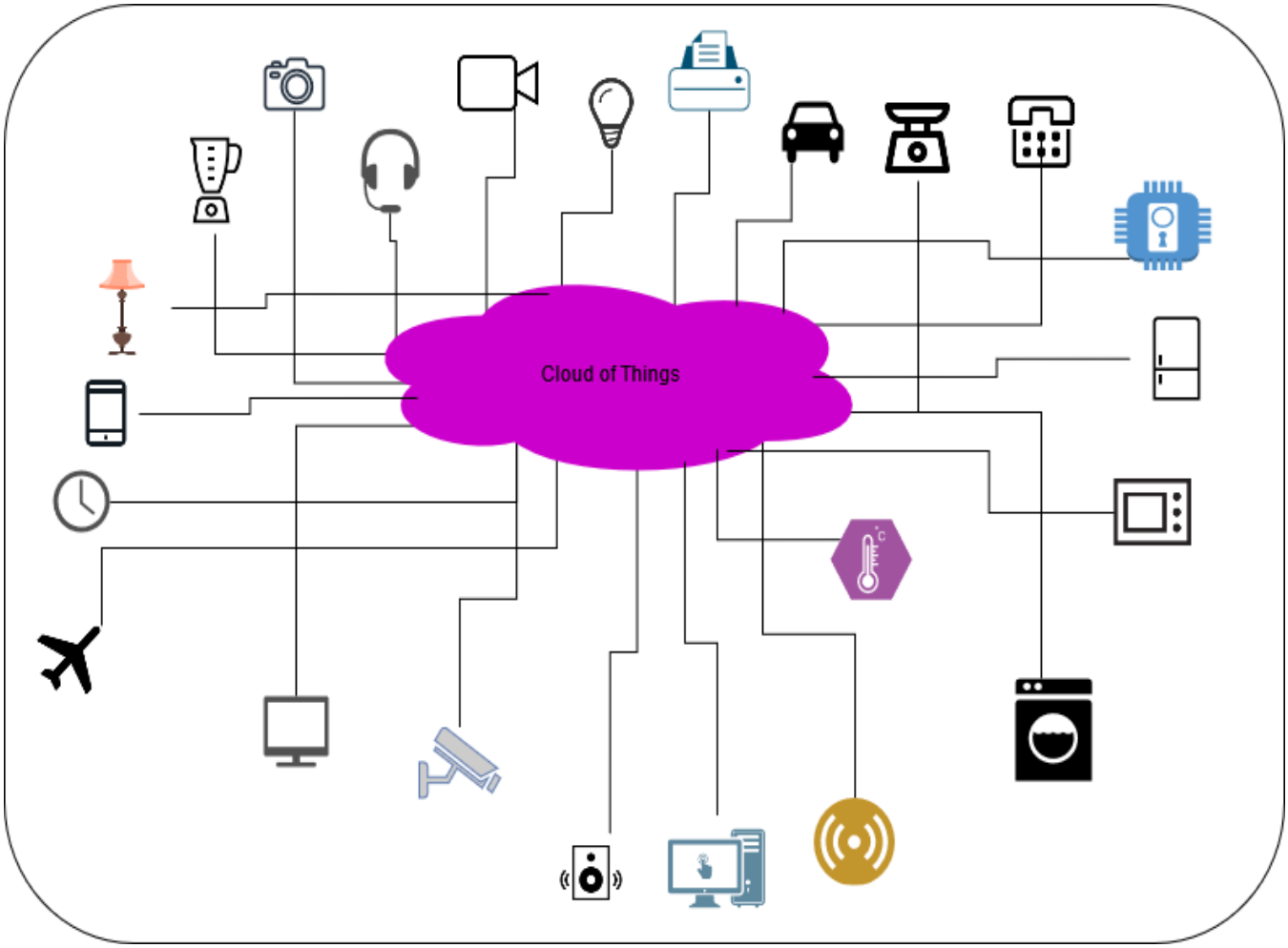


Figure 1: IoT–Connected Devices and Their Relationship with the Cloud of Things (Ashari, 2021)

A second keyword is the ***Cloud of Things*** (CoT) platform. CoT integrates IoT and Cloud Computing (CC) (Aazam et al., 2014).

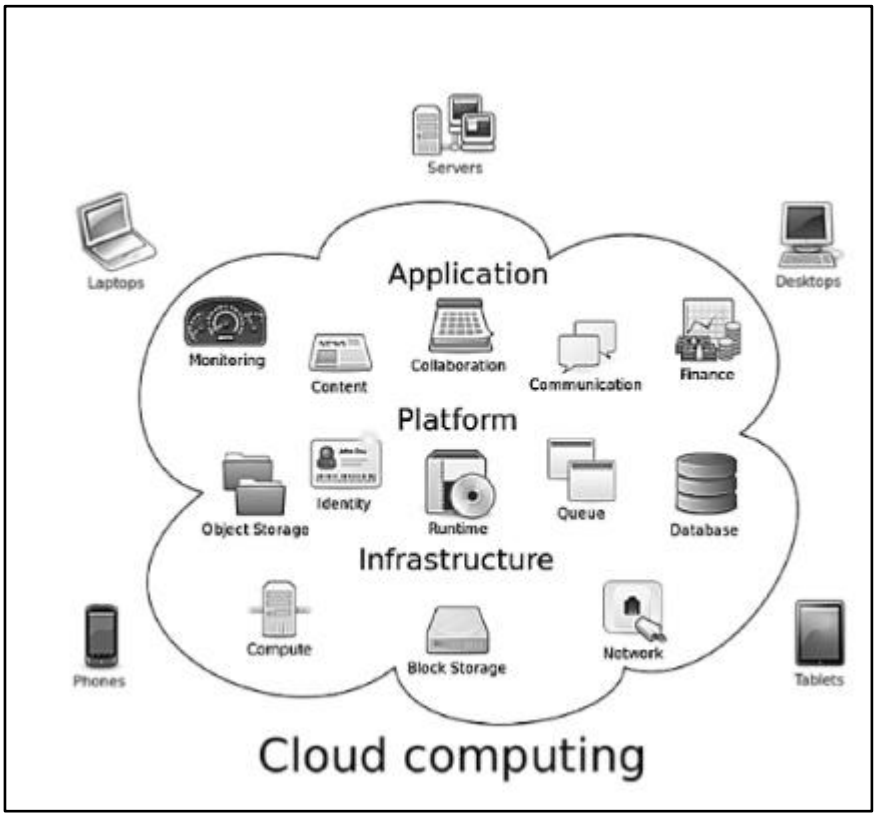


Figure 2: Cloud Computing Architecture (Qu et al., 2017)

While cloud computing offers limitless capabilities to support IoT services and applications by exploiting the data gathered from IoT devices, CoT gives birth to new types of intelligent services, such as video-surveillance-as-a-service, sensor-as-a-service, and big-data-analytics-as-a-service (Sahmim et al., 2017). CoT also provides the required ecosystem on a large scale to deal with multiple Big Data analytics requirements related to mining the data from IoT devices and sensors (Alhaidari et al., 2020). CoT enables various applications to use intelligent things' functions as a service. Multi-tenancy means sharing smart functions by multiple applications (Kim & Kim, 2015). An IoT tenant is a customer who has one or more IoT devices registered with a CoT provider. A dedicated tenant administration platform is allocated to the tenant to monitor and manage their connected devices' data. While the single tenant in the multi-tenant ecosystem of CoT has their own administration space, this tenant still shares the same CoT resource as other tenants.

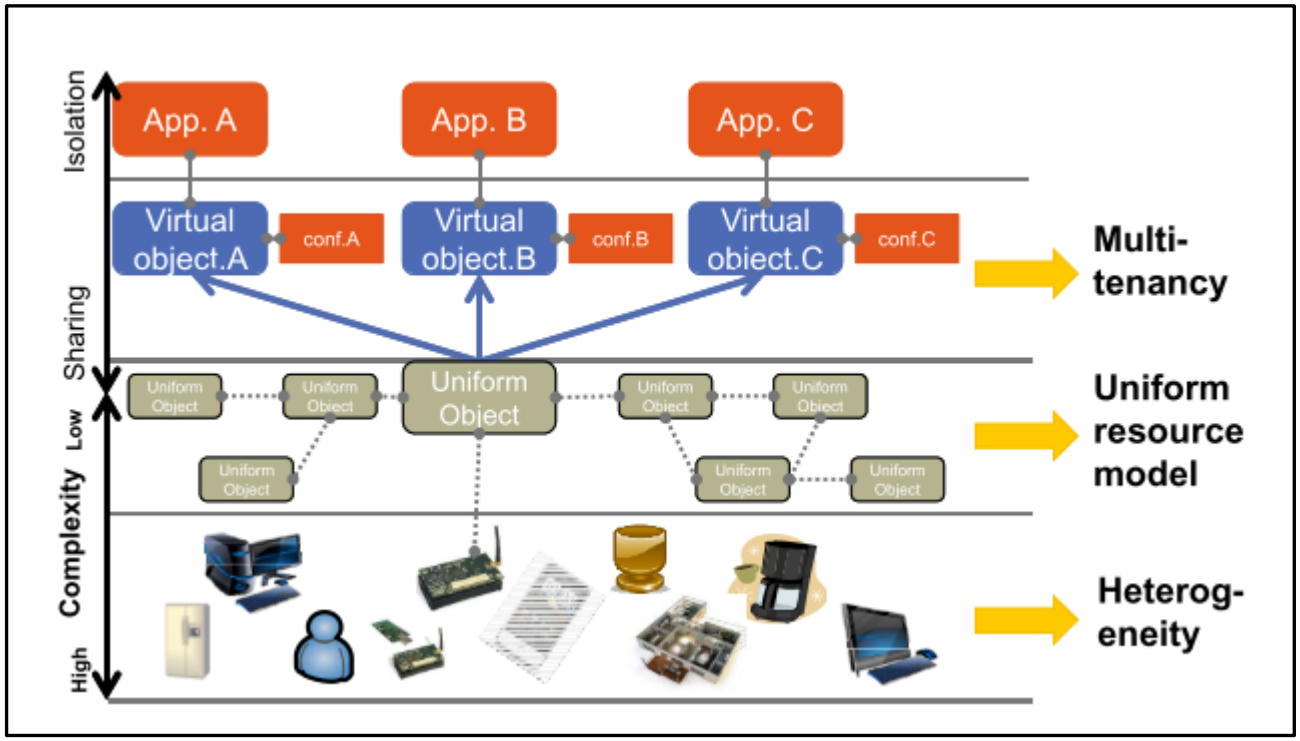


Figure 3: Multi-Tenancy and Sharing of Resources in CoT (Kim & Kim, 2015)

The third important keyword is ***Privacy***. Privacy is having the privilege to be left alone (Lukács, 2017; Warren & Brandeis, 1890). In addition, the right to privacy should include the right to be left alone, limited access to self, secrecy, control over personal information, personal protection of identity and dignity, and finally, intimacy (Solove, 2008).

Related to privacy, another important keyword is ***Natural Person***. The General Data Protection Regulation (GDPR) Article 4(1) specifies that privacy only applies to natural persons, also known as ***Data Subject***, whose data is being collected, stored, and used (Cerbo et al., 2018).

Article 4(1) of the GDPR defines ***Personal Data*** as "any information relating to an identified or identifiable natural person (data subject)." An identifiable data subject is a person that is recognisable due to their personal data information or the linkable information that can single out this person from other people (Cerbo et al., 2018).

Closely related to Personal Data are the keywords ***Data Processor*** and ***Data Controller***. GDPR Article 4(8) defines Data Processor as a person or company that collects personal data on behalf of the Data Controller and can also be the Data Controller themselves. Data Controller, on the other hand, according to Article 4(7) of the GDPR, is the person or company that decides why and how to process the personal data (Cerbo et al., 2018).

The following terminology is the key to this study. ***IoT devices*** in the connected, networked application can ease the burden of manual monitoring. For example, the environmental data from IoT-managed roads and highways improved routing and navigation, enabling the implementation of autonomous cars and trucks (Manyinka et al., 2015). A pacemaker monitors a patient's health data (Dimitrov, 2016) by giving real-time alerts to their doctors in case of an emergency. Inventory management IoT sensors assist companies in automatically performing analysis on supermarket stock requirements just-in-time (Manyinka et al., 2015). IoT applications could also further

reduce the cost of production of goods and services as the producers use IoT systems to improve product design and operations (Manyinka et al., 2015).

However, using IoT also can expose users to cybercriminals who would like to steal and misuse personal private data. IoT devices monitored via mobile phones can function as hackers' tracking devices to get user data, daily routines, or behaviour patterns (Vojković et al., 2020). The term ***Private Data Leakage*** is also known as Information Leakage or Data Breach (Pigni et al., 2018). This term refers to the coincidental, unwanted, or accidental disclosure of sensitive, confidential, or protected private data without the owner's consent (D'Acquisto et al., 2015).

The thesis, in general, focuses on the applicability of AI techniques for identifying data leaks, from IoT applications to the CoT provider, with the investigated subdomains being machine learning and rule-based methods. With this understanding, the following keyword is ***Machine Learning*** (ML). Machine learning can emulate human intelligence by learning from the surrounding environment (El Naqa et al., 2015). A classical approach to ML takes a massive amount of data, computing certain 'features' and training a classifier using these 'features' to become a 'model.' This study will review ML techniques to identify private data leak events. Apart from ML, the final important keyword is the ***Rule-Based*** method. This method is famously known as the IF-THEN statement, which specifies the condition to be satisfied (the IF part) to lead to an action (the THEN part) (Abraham, 2005).

1.3 Research Question

This study investigates how to detect private data in the incoming stream of an IoT payload by analysing ML and RB techniques. The secondary aim of this study is to analytically compare the observation of detecting personal data between these two primary techniques.

1.4 Project Aims and Objectives

This study will first identify the data types in the IoT payload stream. The study will review a few types of IoT devices and the potential PII data transmitted to the CoT.

This study gets support from Deutsche Telekom (DT), a German telecommunication company, in which IoT and CoT services are one of its various portfolio offerings (Deutsche Telekom, 2021). DT IoT offers its customers various IoT applications and products. For example, IoT sensor connects, IoT goods tracking, IoT process digitalisation, IoT data analysis, IoT hardware, IoT CoT, IoT hub, and IoT Solution Optimizer (Deutsche Telekom, 2021). Our scope for this study will be incoming IoT data streams from IoT devices to the DT CoT.

Figure 4: Deutsche Telekom Cloud of Things (Deutsche Telekom, 2021)

DT introduces its CoT as the high-performance, cloud-based IoT application platform for the IoT. The customer can use the CoT to connect remotely, control, and monitor their IoT–connected devices and machines. The CoT collects and analyses live IoT data from the connected devices and appliances to achieve this automation. This study will investigate what PII data is included in the IoT CoT setup of DT from this point of view. The objective is to confirm the identity of the PII, including the unique characteristics and representation in the IoT payload stream. This study will review the IoT CoT technical documentation from DT thoroughly to verify the types of PII that can leak.

1.5 Research Methodology

The principal methodology for this study is a qualitative/theoretical approach. The study will also gather information on private data based on previous literature on personal data characteristics. The main method is to analyse data based on the IoT payload stream for CoT.

The study will also interview personnel from DT Privacy to discuss the applicability of this research study to the DT business. Additionally, we will interview the DT AI department personnel to validate the ML methodology proposed in this study's analysis and design chapter.

The study will concentrate only on the privacy part of information security. In principle, we cannot discuss privacy without security in mind because these two aspects are interrelated. However, the main idea of the research question is regarding private data leaks in the IoT data stream. This study assumes that the underlying structure of the IoT end-to-end data flow is secured.

1.6 Potential Challenges

This study would like to bridge the gap between IoT and Privacy. Consequently, it needs to find the overlapping factors between these two topics and find a way to identify how to extract private information from the IoT data stream. Identifying personal information in the IoT data stream is not easy since we have too many IoT devices and too much information, including the private data accumulated at the IoT data stream. It is essential to understand the data structure, position of personal data in the data stream, and what kind of PII information is stored in the backend.

1.7 Structure of the Thesis

The thesis begins with an abstract for the research topic. The first chapter introduces the core motivation for selecting the research question, frequently used keywords, research methodology, and the potential challenges and limitations.

The second chapter will reveal the background of the study that is the basis of the research topic. This chapter will also explain which source of study data will be observed and analysed. The second chapter will also describe in detail the unique setup within DT. Relevant literature will be reviewed and linked to the research questions.

The third chapter will discuss the analysis of the private data and the method to analyse the data via ML and RB techniques. This chapter will only focus on the incoming personal data to the DT CoT. The study will compare the advantages and disadvantages of both ML and RB techniques and their limitations and potentials.

The fourth chapter deals with the results and the evaluations of the analysis and design in chapter 3. In addition, this study will add findings from an interview with a DT AI expert related to the machine learning methodology. The DT Privacy expert will give his opinion on the research objective's relevance to the DT IoT business.

The concluding chapter will summarise the entire research analysis, results from interviews data, and the recommendation to DT on which method is better to identify and process private data in an IoT payload stream.

1.8 Chapter Summary

This chapter has introduced why IoT and Privacy are the two main characteristics that inspire the research topic. The research aims to provide a method for an organisation to proactively create an alert system for private data leakage. Here the chosen research methodology and its potential challenges and limitations are also described.

Chapter 2. Background and Review of Literature

2.1 Chapter Introduction

This chapter presents the research questions and literature reviews related to the topic and explains the setup with the Deutsche Telekom Cloud of Things.

2.2 Research Background

The study focuses on the IoT data stream payload from the registered IoT devices into the CoT. With this background, the study will try to identify which part of the IoT data stream belongs to private data with the ML approach and the RB approach.

The ML approach contrasts with a deterministic RB approach. This case study could identify personal information in a document (Cerbo et al., 2018) or a string of IoT payload streams. Cerbo (2018) summarised the advantages of ML as follows:

> ML can increase flexibility in detecting personal data using only a single ML model to see multiple individual data categories without specifying uniquely identified parameters. ML learned which specific parameters to consider relevant for the personal data categories after numerous iterations of going through a huge amount of data.
>
> Further, ML can increase the robustness of the recognition system by detecting a class of information even if misinformation or errors exist, such as typo mistakes or other inconsistencies in the format.
>
> Finally, according to Cerbo 2018, ML can simplify system maintenance because the new category of data can be adapted to the model by re-learning and re-training the system. On the other hand, the RB approach requires modifying a complex RB regular expression every time there is a new set of criteria.

Considering the advantages mentioned by Cerbo (2018) above, ML, with a correct ML model, should be able to identify private data in the IoT data stream, which is something this study will attempt to do.

This study will use the Software AG KNIME platform for an ML approach: an open-sourced data analytics, reporting, and integration platform.[1] We chose the Decision Tree and the Naïve Bayes classifiers in the ML approach to train the ML model for simplicity and linear run-time reasons. As its name suggests, Decision Tree is a tree-like algorithm. The best attribute is chosen at the tree's root node, and child nodes represent each possible value of this selected attribute (Jiang & Li, 2011). Naïve Bayers is a probabilistic classifier that assumes that feature variables are considered independent given a class variable (El Naqa et al., 2015).

[1] https://www.knime.com/

Decision Tree and Naïve Bayes are both supervised machine learning methods. Supervised learning feeds the learning system with the example of input-output pairs, and the goal is for the ML model to learn how to map an input to the output (Zhou et al., 2017).

For the RB method, this study will use regular expression. Regular expression filters or extracts tasks, results, or lists from a large dataset via a carefully constructed algorithm or 'expressions' (Li et al., 2008). The regular expressions represent the rule-based conditions as in IF-THEN statements.

In the next section, this study will explain the source of CoT and IoT payload stream specifications discussed further in chapter 3.

2.3 Special Setup with Deutsche Telekom

This study makes use of the unique setup within the Deutsche Telekom IoT department in terms of access to internal and confidential documents of CoT services as follows:

The System Description document explains the DT CoT setup, core architecture design, and how DT manages its CoT in the multi-tenancy ecosystem (Deutsche Telekom Group, 2020).

The Data Privacy document describes the personal data in the CoT in terms of data types, data categories, and the governance of these data types involving multi-tenancy according to the GDPR (Deutsche Telekom AG, 2021).

The Authorisation Concept document lists the tenants' authentication and authorisation procedures and how the DT administrators can assist in the re-authentication request (Deutsche Telekom Group, 2021).

It is important to note that the DT CoT uses Cumulocity IoT Representational State Transfer (REST) and HTTP Application Programming Interfaces (APIs) (Deutsche Telekom Group, 2020). Cumulocity IoT is an open API specification by Software AG that provides their complete API functionality as an open source (Cumulocity, 2021). Cumulocity IoT APIs utilise resource-oriented URLs. It accepts form-encoded request bodies, acknowledges them with JSON-encoded responses, and uses the standard HTTP/HTTPS protocols (Cumulocity, 2021).

Throughout the study, a dedicated internal T-Systems (a subsidiary of DT) supervisor specialised in the Data Privacy department reviewed and revised the thesis document to align with the objective. The internal supervisor reinforced the thesis's quality and gave feedback from a DT data privacy point of view. The thesis findings will help create an alert event system in case of a data breach in the DT CoT.

The following section will discuss the literature studies on relevant topics that motivated the research questions selected for this study.

2.4 Literature Review

IoT and privacy are two main ideas that motivate the research question. IoT is ubiquitous, and almost no one can avoid using at least one IoT device in their home or personal transportation or in public spaces. One of the most common IoT devices people use is fitness trackers (Dimitrov, 2016), which track users' fitness activities, exercise, and food intake. In transportation, the Global Position System (GPS) enabled navigation systems to track the movement and the position of vehicles. Hence, they can estimate arrival time at the destination and give information regarding traffic throughout the planned journey (Renner, 2021). In public spaces, IoT camera surveillance systems exist everywhere to capture the activity of people (Emami-Naeini et al., 2017). These examples introduce one concern: that the individual's data is captured and monitored willingly or unwillingly by these IoT devices.

With this concern in mind, any IoT applications, products, and devices should incorporate the idea of Privacy-by-Design as part of their default build.

2.4.1 Privacy-by-Design

Ann Cavoukian (2012) first introduced the Privacy-by-Design concept. This idea explains how to include privacy at the initial stage of information system development. The Privacy-by-Design concept recommended seven foundation principles to secure personal data. Privacy as the default setting and privacy embedded into the design are among these principles.

This concept, however, did not specify how to incorporate privacy into an IoT application. Hoepman (2014) came up with eight privacy-by-design approaches with which to align in the early stage of the software lifecycle design. The eight approaches are minimise, hide, separate, aggregate, inform, control, enforce, and demonstrate.

Perera et al. (2016) agreed with Hoepman's (2014) approaches and elaborated on the descriptions in their research. Perera (2016) identified major privacy risks in implementing IoT applications: secondary usage and unauthorised access. Secondary usage refers to using data without the data owner's consent of the initially agreed usage, while unauthorised access meant access to data without proper authorisation. Perera (2016) then developed nine privacy frameworks that can be used as guidelines to evaluate privacy capability and identify the gaps in the IoT providers' existing offerings, taking into account the privacy risks of secondary usage and unauthorised access. Perera (2016) divided IoT application data flow into three sections: IoT devices, middleware, and IoT cloud.

This study will concentrate more on the data flow between the IoT devices up to the area where cloud infrastructure (in our case, CoT) resides. Moreover, the study will deal with privacy risks by detecting private data information in the incoming IoT data flow.

Like Perera (2016), D'Acquisto et al. (2015) also concurred with Hoepman's (2014) findings. D'Acquisto (2015) looked at privacy-by-design as a process of implementing privacy and data protection principles in various

technological and organizational components. By incorporating privacy measures and privacy-enhancing technologies (PETs) directly into the design of information technologies and systems, one can bridge the gap between the legal framework and the available technologies (D'Acquisto et al., 2015).

The study will show that detecting a private data leak event in the incoming IoT payload is possible as part of the privacy-by-design measurement.

In a different view, Barr-Kumarakulasinghe et al. (2021) proposed using an IoT regulatory framework to safeguard the private data of the IoT users. More and more privacy-by-design applications of IoT technology are emerging in the market, and this factor increases end users' confidence in adopting IoT products, applications and services. The IoT regulatory framework can act as the check-and-balance of this phenomenon (Barr-Kumarakulasinghe et al., 2021).

The next literature reference related to privacy-by-design is from Hintze et al. (2017). Here, privacy-by-design was seen from a different angle on the controlled linkable data. Hintze (2017) proposed that the organisation decouple data elements from re-identifiable linkages to data subjects or owners while maintaining the applicability of the data to be further analysed for its consented purpose. It enables data controllers or data processors to meet the precise exclusionary standards of GDPR Articles 11(2) and 12(2) on the controller, demonstrating that they are not in a position to identify the data subject.[2]

Although removing the linkable identifier to the IoT data represents the next step and is not covered by this study, it is part of the complete pipeline towards enhancing the privacy-by-design solution for DT IoT in the CoT. The first step is identifying the private data and labelling it as personally identifiable information (PII).

IoT devices, as mentioned earlier, transmit a massive amount of data to the CoT. This data, also known as big data, has a lot of information to analyse and mine. Hence, the same concern arises: whether one can sufficiently protect individuals' private data.

2.4.2 Privacy, Big Data, and IoT

The European Union Agency published a similar topic on "Privacy-by-design in big data" for European Union Agency For Network And Information Security (ENISA) in 2015. This 2015 publication explores the privacy topic thoroughly, introducing the concept of "big data versus privacy" to "big data with privacy." The idea of big data is described in detail and specified in terms of volume, velocity, and variety.

Barnaghi et al. (2013) discussed the key issue in big data collections: quality, validity, and trust when multiple parties accessed the collected data. In a similar view, ENISA listed seven challenging points of big data collections: lack of control, transparency, data reusability, data inference, data re-identification, profiling, and automated decision making (D'Acquisto et al., 2015). In the example of parking prediction, the accuracy of the available parking prediction will depend on the data analysis from big data collection regarding historical data of

[2] https://gdpr-text.com/

parking, traffic, environmental data, and events of people (Badii et al., 2020). Multiple data sources in this case could lead to privacy risks due to the data owner's lack of personal information control.

Dimitrov (2016) expressed his view on big data collection in the medical Internet of Things (mIoT) applicability. According to Dimitrov (2016), communication was one of the major challenges to implementing the mIoT mainly because IoT devices with different manufacturers have different proprietary protocols. Creating standard or globalised protocols for IoT device communication was critical to transforming the collected healthcare big data into meaningful information ready to be used by health professionals in their fields.

In this thesis study, the CoT platform is based on open-source Cumulocity IoT API. Several IoT communications protocols are used and described further in chapter 3 in the CoT architecture subsection. The data source for our machine learning and rule-based scenario comes from randomly generated JSON files, mimicking the real-life IoT data payload from multiple IoT devices. We eliminate the different communication protocols issue because only Cumulocity API standards are applied and followed.

As presented earlier, there is a concern about protecting private data from secondary use or unauthorised access with IoT and big data. GDPR regulates how organisations can collect and process data and, at the same time, protect personal and private data (Kaneen & Petrakis, 2020). Kaneen & Petrakis (2020) devised a methodology to check GDPR compliance when managing personal data digitally. A remote patient monitoring system was selected to prove this methodology. The IoT device in this case was a heart-rate monitoring sensor that was continuously sending data to an interface in the cloud. Kaneen & Petrakis (2020) tested GDPR compliance for the relationship among the device interface, web application, and application logic. Moreover, Kaneen & Petrakis (2020) developed a platform-independent query language for evaluating GDPR compliance. The compliance decision is based on the thorough observation of the inspected system and the movement of personal data until the end of the queries chain.

Badii et al. (2020) commented that in terms of GDPR compliance, in the context of IoT, different solutions may need to cater to additional requirements depending on the IoT application's purpose. The healthcare IoT application requires strict protection of personal data, a contrast to the entertainment IoT application such as virtual reality video games (Badii et al., 2020). Every IoT application introduces a potential threat for private data leaks. Companies need to balance the risk of private data leaks, compliance with GDPR, and their business objectives.

Considering GDPR compliance, this study should understand which threat models to apply in our IoT data leak scenario. We describe the IoT threat model in chapter 3.

2.4.3 IoT and Artificial Intelligence

This literature review topic can show the relationship between IoT and artificial intelligence (AI). IoT, as mentioned previously, always deals with big data and functions to automate certain procedures via actuators or reporting events and create alarms via the IoT sensors. On the other hand, AI is a special branch of computing that deals with predictive analysis and automated decision-making (Galaz et al., 2021). Kishor & Chakraborty (2021) viewed IoT as a catalyst that amplifies AI applications' potential. Further, ML and RB are part of AI and can have their own advantages and disadvantages.

When it comes to ML's advantages, Makkar et al. (2021) saw ML as a useful tool to forecast and detect vulnerabilities in IoT-based systems. An ML algorithm can be trained to learn the trends and historical relationships among data (Makkar et al., 2021). Our study will indicate that the ML method can identify private data using supervised learning.

Cerbo et al. (2018) believed that ML was a suitable tool for identifying semi-structured or unstructured data in large repositories or databases. In contrast with the RB method, the ML classifier with the correct ML model could handle the semi-structured or unstructured data. Cerbo et al. (2018) proposed a supervised ML system to identify personal information in the large dataset. According to Cerbo et al. (2018), a Natural Language Processing (NLP) method with predictive classifiers such as Naïve Bayers or Convolutional Neural Network (CNN) could be trained to predict text processing in the unstructured dataset. Topaz et al. (2019) aligned with Cerbo et al.'s (2018) opinion in developing an intuitive approach to extract meaningful information from big data. Our study will also use the NLP approach using Naïve Bayers and Decision Tree classifiers to identify private data information from the class of labelled documents.

Another advantage of ML is related to privacy concerns. Jeong et al. (2017) believed that ML could reduce privacy concerns. In IoT applications where all data travels to the cloud analytic platform, there is a general concern that the servers or clouds platform could easily collect sensitive private information (Jeong et al., 2017). Waheed et al. (2020) perceived that in case IoT exists in numerous important applications, it is as if everyone's data are out on the internet. Therefore, they are more exposed to personal data breaches and identity theft (Waheed et al., 2020). This study addresses this private data leak concerns via the attempt to identify sensitive information from the incoming IoT data payload.

Jeong et al. (2017) proposed in their study a workflow to execute privacy-preserving transmission using a Neural Network (NN) ML algorithm. The privacy-preserving feature occurred at the client side (IoT devices) and partially at the server level. In their study, the IoT devices' system performance depended on which network the end users used when transmitting the data to the cloud. Similarly, Dimitrov (2016) questioned whether executing the ML algorithm would impact the real-time requirement of the IoT application. Although the findings from Jeong (2017) are interesting, our study is not about detecting private data leaks at the client level but instead focuses on the IoT data coming into the CoT platform.

2.4.4 Privacy Breaches and GDPR Fines

The final aspect of the literature review regards privacy breaches and fines. GDPR penalties are one of the primary motivating factors for DT to identify data breaches in their IoT CoT platform. GDPR.eu 2021[3] explained two tiers of GDPR fines: up to €10 million or up to 2% of the organisation's global annual revenue. One factor determining the fine is the violation of specific GDPR articles, for instance, governing controllers and processors (Article 8, 11, 25–39, 42, and 43). Moreover, there are ten criteria to further assist the evaluator in which penalty to give and of what amount.

As a data processor, DT is responsible for analysing the private data according to the purpose of the IoT devices and applications. Further, as the IT provider of CoT, DT provides a platform for its customers to perform data mining and analytics for their IoT-connection applications. With this responsibility, it is imperative to know the consequences of a data breach as specified by GDPR. The research topic is to identify private data leaking, which aligns with the strategy to avoid non-compliance by creating private data events for DT organisational proactive measures.

2.5 Chapter Summary

In this chapter, we discussed the background of the research topic. Then, this chapter explained the relationship with the Deutsche Telekom CoT team and its unique setup. This study chose privacy-related topics as the literature review and the motivating factor for selecting the research question. The next chapter will start with the high-level architecture analysis and design for ML and RB approaches to identify private data in the IoT payload.

[3] https://gdpr.eu/fines/

Chapter 3. Analysis and Design

3.1 Chapter Introduction

This chapter presents the types of PII data and their descriptions. It also explains what kind of IoT devices can transmit PII data and for what purpose. Furthermore, we will present a threat model to describe what kind of threat this study tries to solve against the misuse of private data in the CoT scenario.

Finally, we will explore the capability of the ML and RB as tools for detecting private data (PII) from the IoT payload.

3.2 Personally Identifiable Information (PII)

Any PII for a natural person can disclose their identity, and in the worst-case scenario, it will lead to cyber harm (van den Berg et al., 2014). PII can be categorised as follows (Emami-Naeini et al., 2017; Malandrino et al., 2013; McCallister et al., 2010; Sahmim et al., 2017):

No	PII	Description
1	Information regarding name	This information refers to the data owner's name, parent's name, and maiden name.
2	Personal Identification Number	Social security number/taxpayer identification number like SSN/BSN, Passport number, driver's license number, bank account, credit card number
3	Address information	Address, email address, zip code
4	Personal characteristics	Photographic image (esp. face or other identifying characteristics), fingerprints, handwriting, biometrics data (retina scan, voice signature, facial geometry)
5	Information linked to an individual in one of the PII factors	Information related to birth, religion, physical description, frequent activities, frequently visited places, employment information, medical information, education information, financial information, GPS data
6	Family	Data owner's family members information
7	Phone number	Data owner's telephone number, mobile number
8	Password	Data owner's password
9	Sensitive information	Sensitive private data: memberships, demography data, Internet habits
10	Hardware ID	MAC address, IP address, uniquely identifying the device used

Table 1: PII Categories in General (Ashari, 2021)

After knowing the types of PII, we should also know which kind of IoT devices can transmit this PII information to the internet.

Type of IoT Device	Description	Potential PII
Sensors: Presence	User location is tracked (D'Acquisto et al., 2015) User email address (Hon et al., 2011) MAC Address (Yun et al., 2016) IP addresses (Hon et al., 2011)	PII linkable information Address information Hardware ID
Actuators	Smart actuators to control and regulate temperature and devices (D'Acquisto et al., 2015)	PII linkable information Address information Hardware ID

Type of IoT Device	Description	Potential PII
Machine-to-Machine communication device (M2M)	Examples: intelligent transportation systems (ITS), logistics and supply chain management, smart metering, e-healthcare, video surveillance camera and security, smart cities, and home automation (Mehmood et al., 2015)	Personal characteristics PII linkable information Address information Hardware ID Sensitive information
Fitness sensors	Collect data on the way we walk, sleeping patterns, or other streams of data (D'Acquisto et al., 2015)	Address information Sensitive information PII linkable information
Health-related sensors	Monitor patient vital signs (D'Acquisto et al., 2015)	PII linkable information
Biometrics sensors	Biometric data (Hon et al., 2011)	Personal characteristics

Table 2: Device Types and PII in General (Ashari, 2021)

Next, the study will look at the CoT architecture to understand the end-to-end data flow of the IoT device message transmissions from the end-user domain up to the CoT backend.

3.3 CoT Architecture Design

The following figure shows the CoT architecture design.

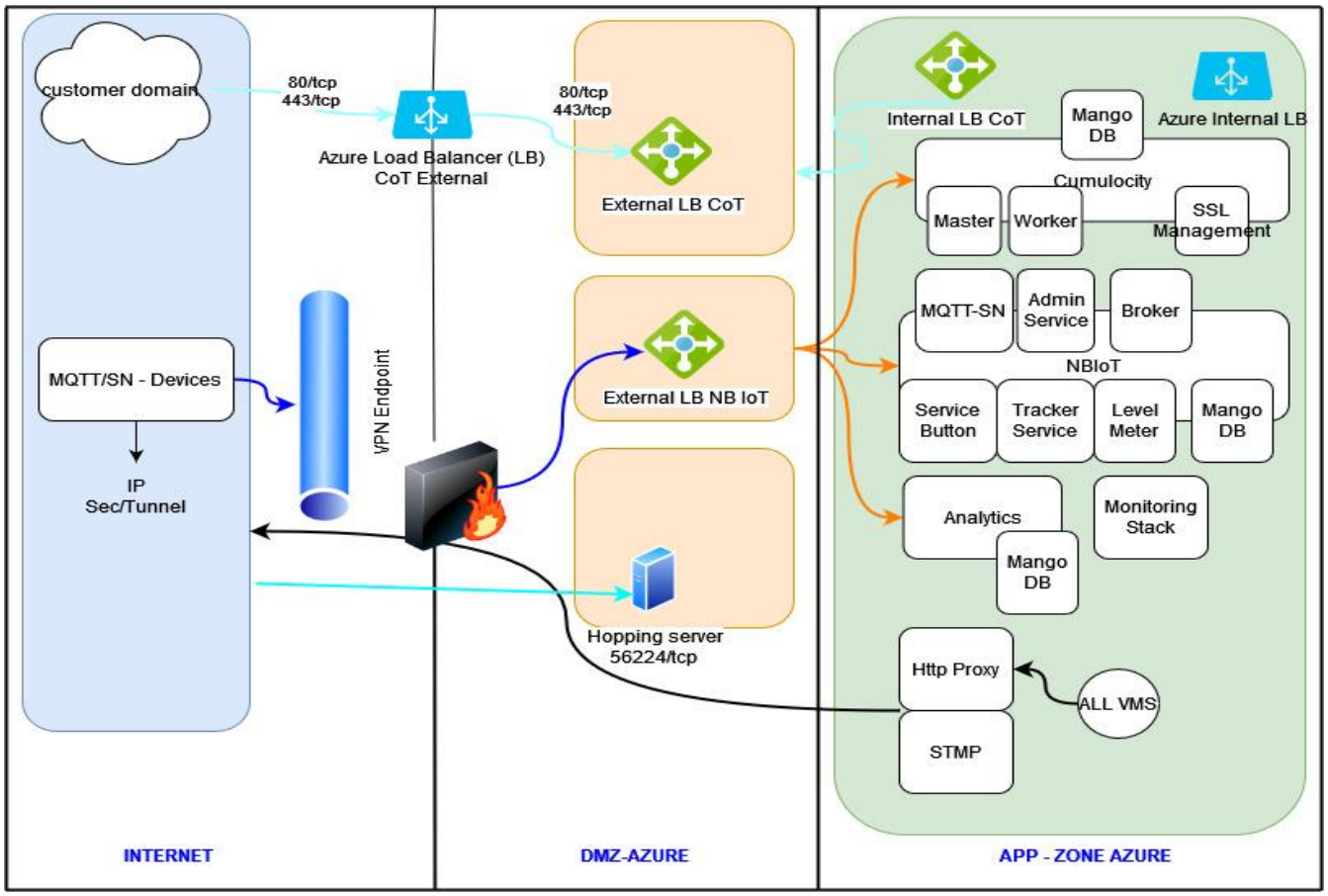


Figure 5: DT CoT Platform Architecture (Deutsche Telekom, 2021)

The CoT architecture design shows how the IoT traffic flows among the IoT devices up to the Azure Application Zone (App-Zone Azure).

For IoT devices with Message Query Telemetry Transport (MQTT) for Sensor Networks (MQTT-SN) protocol, the traffic will go through the virtual private network (VPN) endpoint to check their access after going through the same Internet firewall to the demilitarized zone (DMZ). MQTT-SN is an optimized version of the IoT communications protocol[4]. MQTT is explicitly designed for efficient operation in large low-power IoT sensor networks (Fathy et al., 2018). The traffic will connect to the external load balancer (LB) Narrow Band (NB) IoT from the DMZ. NB-IoT is a massive Low Power Wide Area (LPWA) technology proposed by the 3rd Generation Partnership Project (3GPP) for data perception and acquisition intended for intelligent low-data-rate applications, such as smart metering and smart environment monitoring (Chen et al., 2017). According to their domain model types, the IoT data finally arrives at the NB-IoT section in the application zone.

While the IoT devices potentially transmit various types of PII, the DT CoT has a specific PII registered in the IoT backend. Table 3 illustrates registered data types in the CoT backend during the user's IoT device registration and their purposes.

Potential PII Information	Used for	Privilege
Username	User login	Username can be changed by the user/tenant, with user management privilege
Password	Initial authentication	A user enters a password into the CoT user interface (UI)
Email address	Initial authentication	A user enters an email address into the CoT user interface (UI)
Mobile number	Two-Factor authentication	The user registered mobile number receives a token number via SMS

Table 3: PII Information in the CoT Backend (Deutsche Telekom Group, 2021)

The customers/tenants might have more than one IoT device to replace or change. Despite this, all IoT devices are uniquely identifiable via the MAC address. The object identification procedure specified that each managed object has its unique generated global identifier (Cumolocity, 2021). The identifier is persistent, regardless of the re-purpose of the functionality or replacement of IoT devices.

As table 3 specifies, at the CoT backend, each IoT device is registered with the backend with its PII information during the device setup. Due to this factor, the collected PII information has sensitive data label classification, and DT has built access rights management to restrict access to this sensitive information (Deutsche Telekom Group, 2021).

There are four aspects of IoT sensor capabilities as defined by Cumulocity 2021. Cumulocity defines its sensor library according to the measurement and control capabilities. It pre-packaged the sensor library into four domains: environment capabilities, energy, location capabilities, and common capabilities (Cumulocity, 2021).

[4] https://ambisecure.ambimat.com/mqtt-sn-lowering-the-cost-of-iot-at-scale/

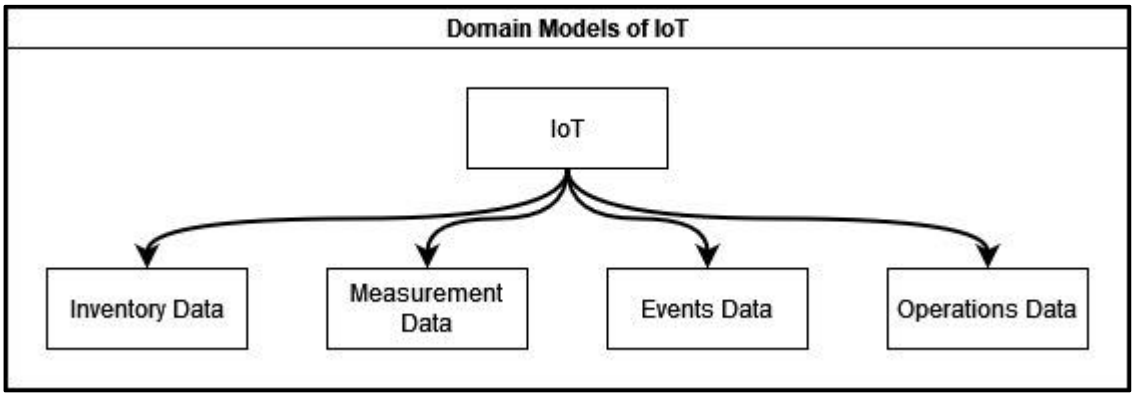


Figure 6: IoT Domain Models (Cumulocity, 2021)

Within these four aspects, the study will examine the linkable information that will recognize a natural person from the incoming IoT data to the front end of the CoT platform.

Knowing the CoT architecture, one can build the threat model to visualize the focus of this study and how personal data identification could help secure private data in the IoT data flow.

3.4 Threat Model

IoT data transmit PII and non-PII data from an untrusted network to the CoT network with the frequency specified during the device registration step. The PII data sent by the IoT devices in the form of a comma-separated-values (CSV) file, also known as IoT data payload, will be analysed by the analytic engine in the Cumulocity CoT. As described in the previous section and illustrated architecture diagram of CoT, the result of this analysis will be re-transmitted back to the IoT devices' owner and give the services or functions they require, for example, whether to switch on or off the lights or increase or decrease the room temperature.

Figure 7 shows the current data flow among different types of networks, i.e., internet, DMZ, and CoT network, assuming the network is secured. We have data owners of the IoT devices on the untrusted network and the DT data processor/controller on the CoT network. The existing private data controls are in place, marked as C01, C02, and C03.

Here we can also see the private data breach threat model that specifies private data threats actors. This study introduces another private data control in the PII detection tool (C04).

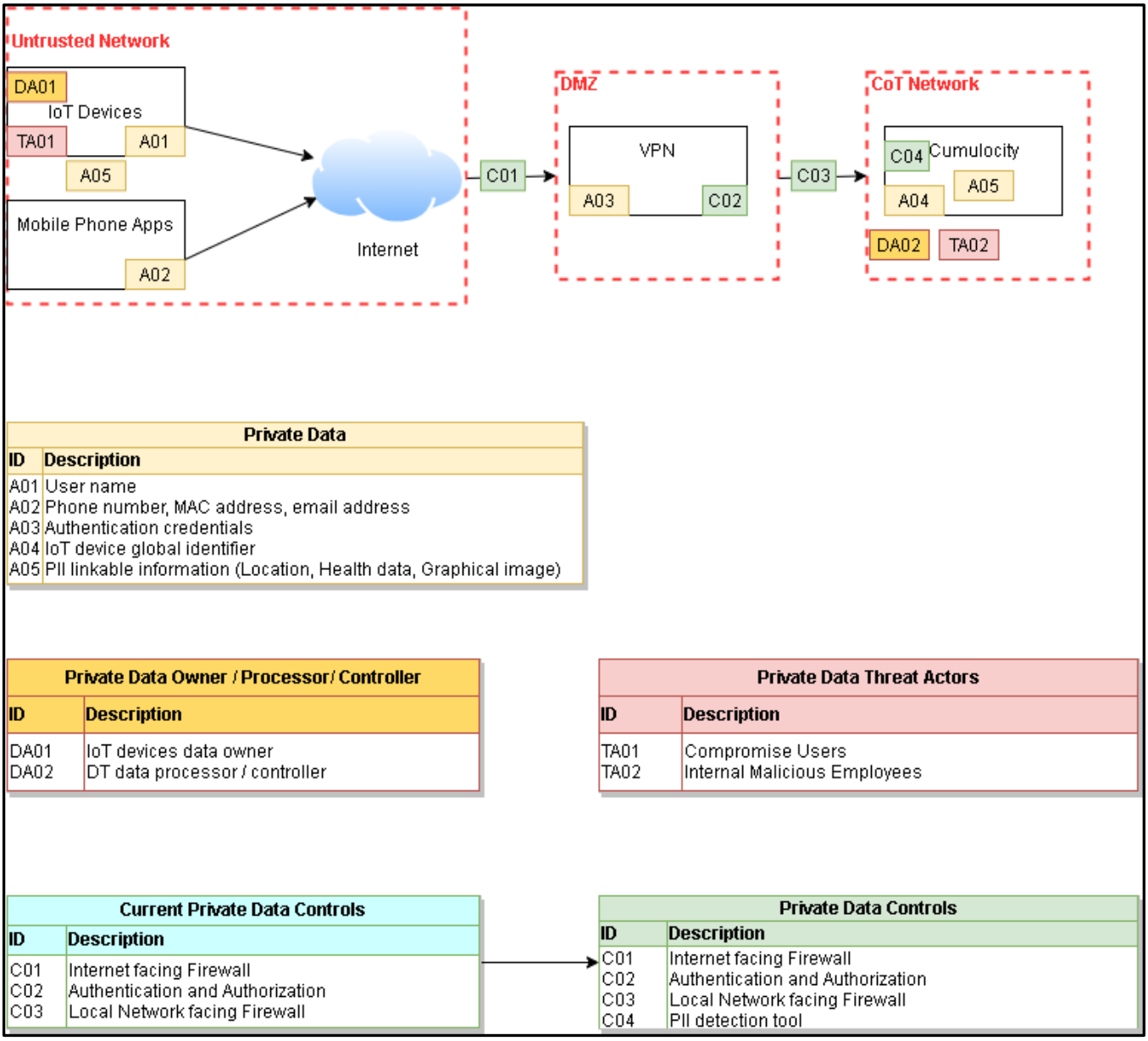


Figure 7: CoT Threat Model (Ashari, 2021)

There are two types of identified threat actors in this scenario. First, we have a malicious user disguised as the owner of the IoT devices or the IoT device application from the mobile phone. This user can send bogus data with IoT payload to the CoT network's analytic section with legitimate authentication. The motivation to do this is to flood the traffic of the analytic section of CoT with a lot of tampered IoT data, including PII. Consequently, this action would paralyse the analytics engine, the same as the function of a denial of service (DOS) attack by the cybercriminal on a particular target network. In this case, it is vital to verify the origin and integrity of the IoT data before it is allowed to enter the application zone network.

The internal threat actor in the threat model is an internal employee. Internal employees have access to the user management tools based on the roles granted to them. Disgruntled employees can harm private data, including PII, and collaborate with external threat actors (Abomhara et al., 2015). This study aims to protect the PII against the second threat actor scenario and prevent DT from collecting more data than needed, leading to a possible

GDPR compliance issue. The PII tool presented in this thesis analysis will tag the PII data, observe their behaviour, and raise a flag in case unauthorised changes happen.

We know which threat actors this study would like to protect the private data from by analysing the threat model. The next section describes the data modelling within the CoT. This study will narrow down to only one type of sensor library as defined by the Cumulocity IoT and apply the method of ML and RB to identify and classify the PII.

3.5 Data Modelling and Classification

The IoT device registers a template with SmartREST proxy in the Cumulocity IoT and sends their stream of IoT messages. The IoT data stream consists of strings of characters in comma-separated-value (CSV) format. SmartREST API converts the message into a corresponding REST API call. In Figure 8, we give an example of the incoming IoT data payload header to SmartREST API.

```
POST /s HTTP/1.0
Authorization: Basic ...
X-Id: ...
Transfer-Encoding: chunked

100,1234456
```

Figure 8: Example of the Incoming Data Payload to SmartREST (Cumulocity, 2021)

Within this data stream, we can identify the personal data based on the data type of the PII described in table 1. Figure 9 shows the observation point of the data stream from the architectural point of view.

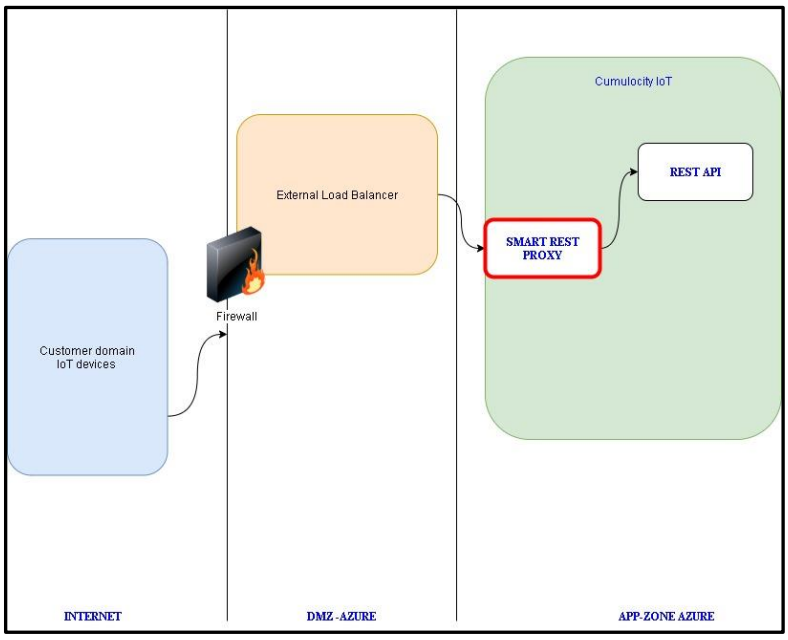


Figure 9: IoT Data Stream to SmartREST (Ashari, 2021)

One of the most linkable PII data of a person is their GPS location. The cybercriminal can detect the potential target location when the unique device ID that collects GPS data is exposed or tracked. In an IoT data payload, the specific fragment name API format called c8y_Position defines the location with altitude, longitude, and latitude (Cumolocity, 2021). We can see the example of an extract from a JSON file with this format in Figure 10.

```
"c8y_Position": {
  "alt": 67,
  "lng": 6.15173,
  "lat": 51.211977,
  "trackingProtocol" : "TELIC",
  "reportReason" : "Time Event"
}
```

Figure 10: Example IoT Location Data (Cumulocity, 2021)

In the following ML analysis, this study analyses only one fragment name, cy8_Postion, as the example of PII in the IoT data payload. This analysis aims to identify PII (cy8_Position) among all other fragment names in the CSV files that arrive at the smartAPI engine.

This study utilises KNIME,[5] the open-source data analytics, reporting, and integration platform for the ML approach. Using the text processing functionality in KNIME, as shown in Figure 11, the study builds an ML pipeline to identify the location data and classify it as PII. The pipeline uses supervised learning to train the ML model.

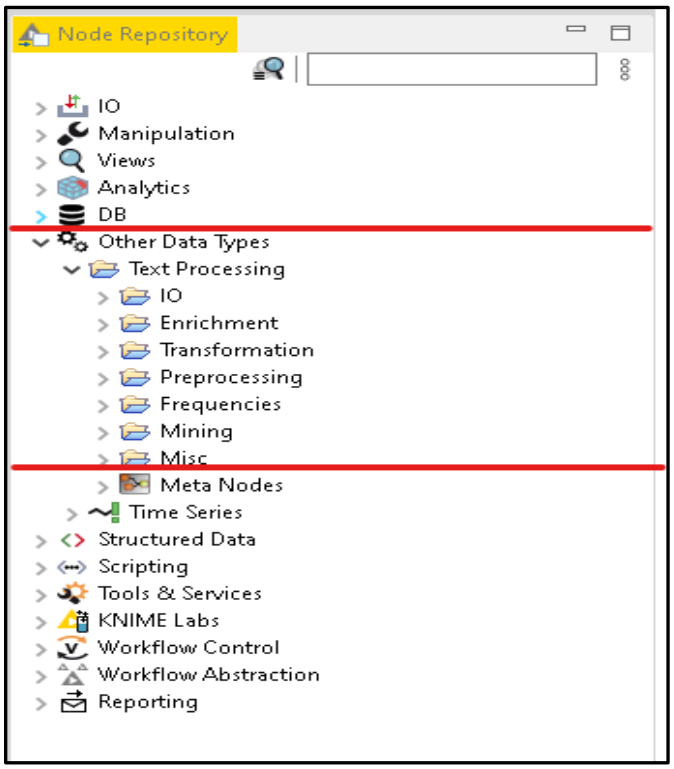


Figure 11: KNIME Node Repository (Ashari, R. 2021)

[5] https://www.knime.com/

Figure 12 shows the process of building a pipeline in KNIME. An ML pipeline is a mechanism to organise data workflow input and output from an ML model (or set of multiple models). It includes feeding the raw data input, preprocessing, transforming and manipulating data output, training and creating a learning model, and finally, giving the prediction output (Zhou et al., 2017).

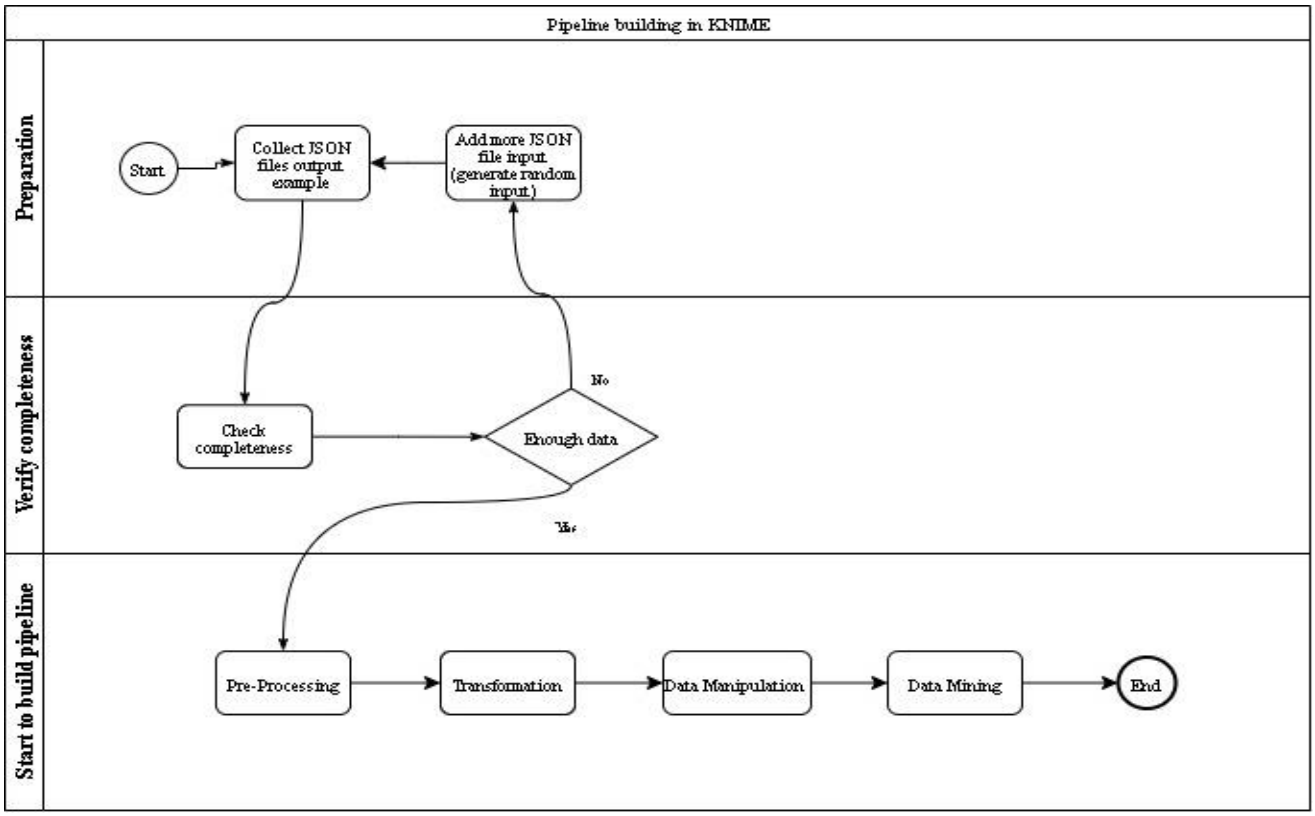


Figure 12: Flowchart of KNIME Pipeline Activities (Ashari, 2021)

In the data preparation stage, using a text file as an input, the data is fed into the file reader in KNIME. All fragment names examples in the Cumulocity sensor library are put together in the text file on the preparation side. The file reader node reads the input text file with JSON file sensor capabilities samples.

The steps to build the pipeline are as follows:

Using a Python script, we generate two folders of JSON files containing strings of fragment names and their corresponding content. The content of the files represents the IoT payload after the conversion to REST API formatting exactly in the same way the content fragment names are used in the Cumulocity document to describe the real data stream.

One folder contains JSON files with all fragment names except cy8_Position; in this case, it is a negative scenario. Another folder contains JSON files with all fragment names, including cy8_Position, and it is a positive scenario.

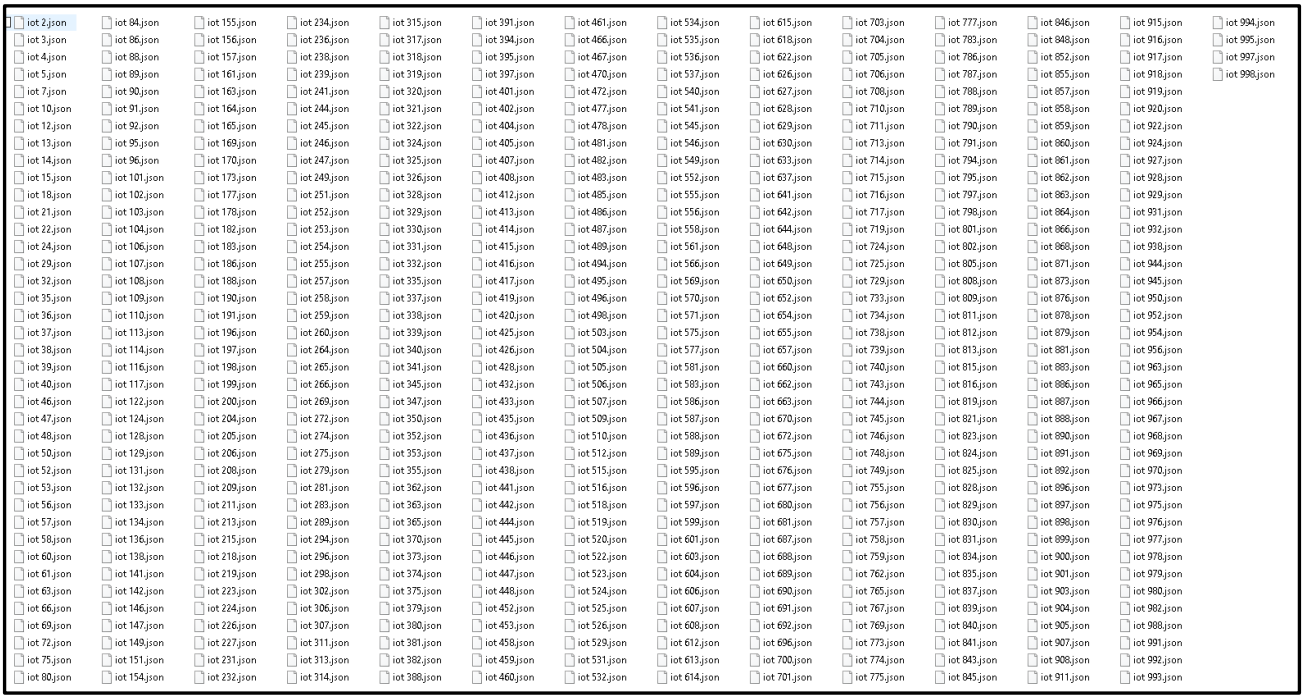


Figure 13: Randomly Generated JSON Files (Ashari, 2021)

The flat file document parser nodes fed the pipeline's workflow using a supervised learning approach, with JSON positive and negative scenarios files. Both the positive and negative file scenarios are combined as an output of the input-output (IO) phase through the concatenate node.

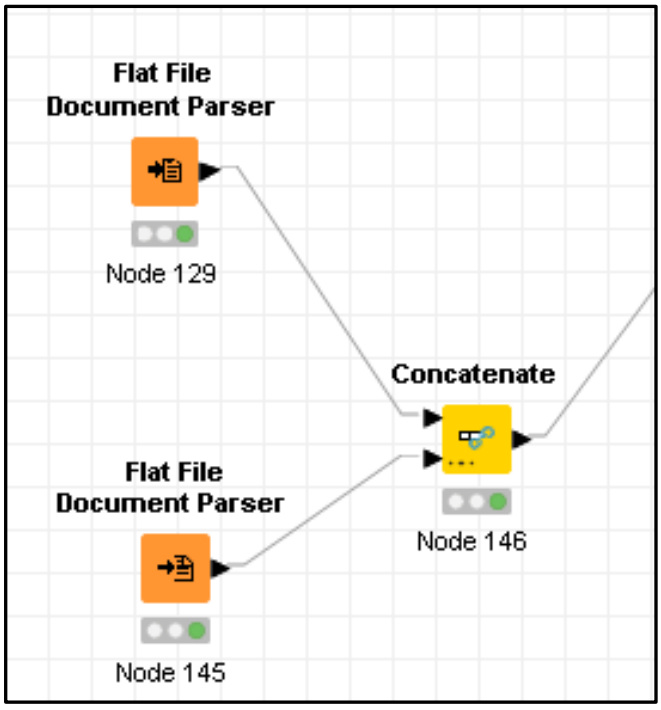


Figure 14: Input to the Pipeline (Ashari, 2021)

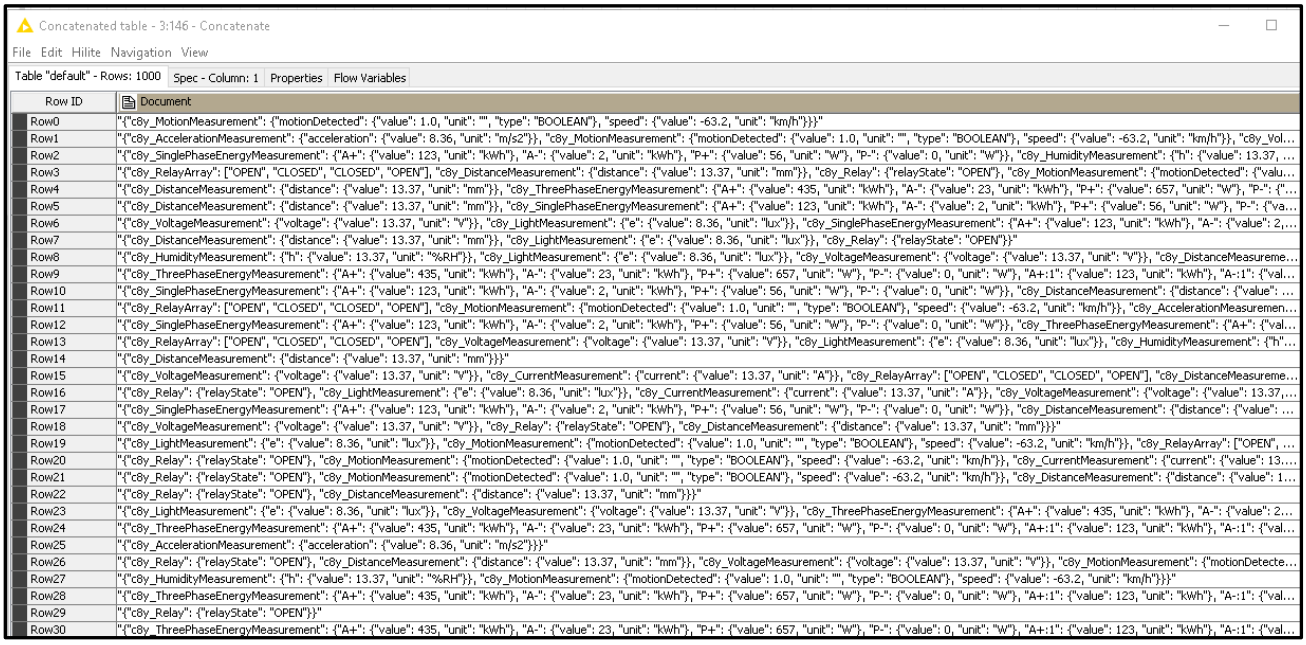


Figure 15: Output from Concatenate Node (Ashari, 2021)

The next stage is to preprocess the data. The punctuation erasure node removes all the punctuation characters of terms in the input documents (KNIME, 2021). The number filter node removes the numbers in the input document.

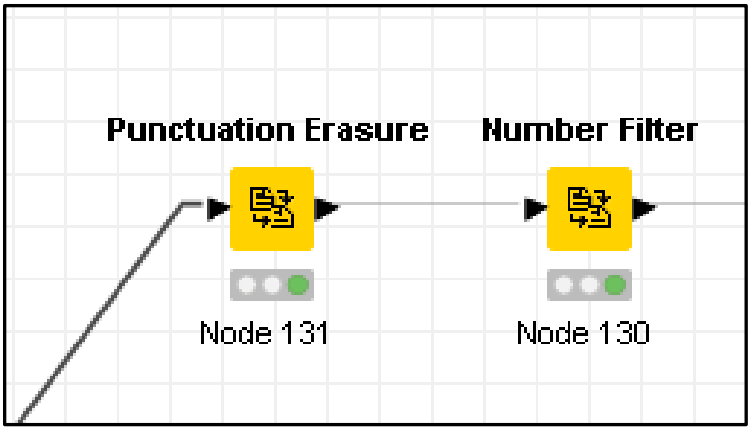
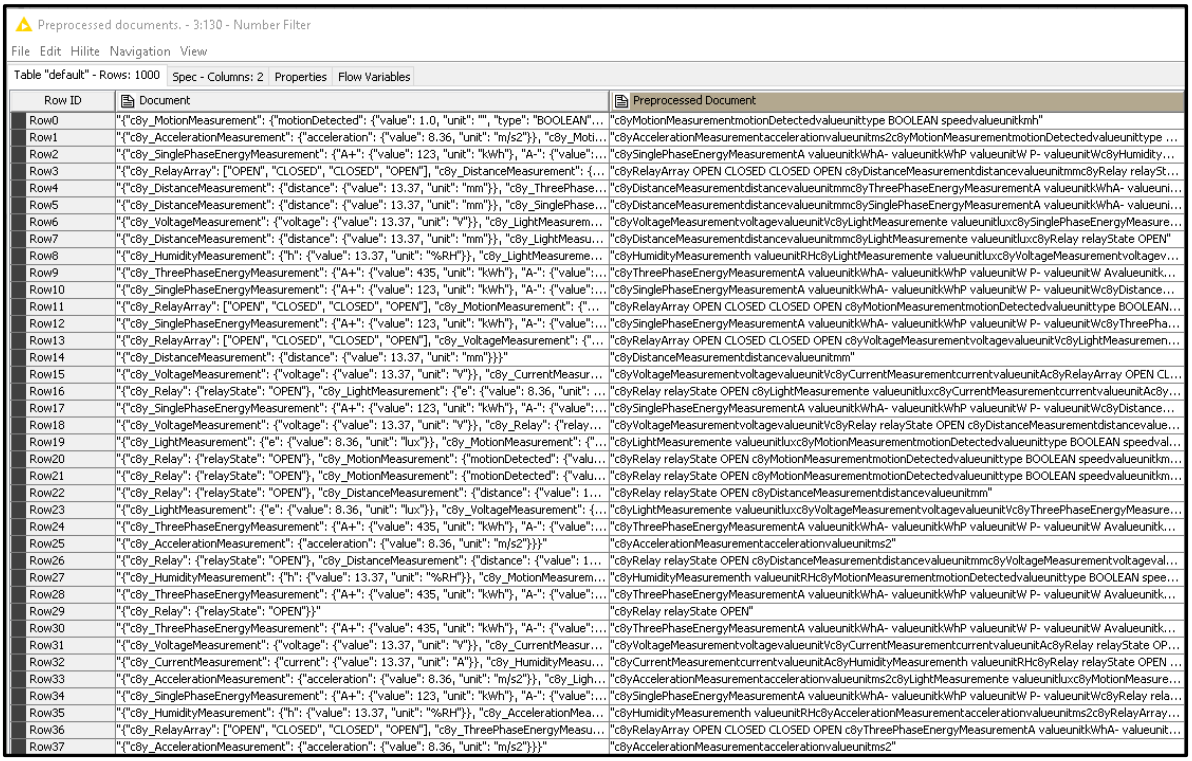


Figure 16: Preprocessing Input Data (Ashari, 2021)



Preprocessed documents. - 3:130 - Number Filter

File Edit Hilite Navigation View

Table "default" - Rows: 1000 | Spec - Columns: 2 | Properties | Flow Variables

Row ID	Document	Preprocessed Document
Row0	"{"c8y_MotionMeasurement": {"motionDetected": {"value": 1.0, "unit": "", "type": "BOOLEAN"...	"c8yMotionMeasurementmotionDetectedvalueunittype BOOLEAN speedvalueunitkmh"
Row1	"{"c8y_AccelerationMeasurement": {"acceleration": {"value": 8.36, "unit": "m/s2"}}, "c8y_Moti...	"c8yAccelerationMeasurementaccelerationvalueunitms2c8yMotionMeasurementmotionDetectedvalueunittype ...
Row2	"{"c8y_SinglePhaseEnergyMeasurement": {"A+": {"value": 123, "unit": "kWh"}, "A-": {"value":...	"c8ySinglePhaseEnergyMeasurementA valueunitkWhA- valueunitkWhP valueunitW P- valueunitWc8yHumidity...
Row3	"{"c8y_RelayArray": ["OPEN", "CLOSED", "CLOSED", "OPEN"], "c8y_DistanceMeasurement": {...	"c8yRelayArray OPEN CLOSED CLOSED OPEN c8yDistanceMeasurementdistancevalueunitmmc8yRelay relaySt...
Row4	"{"c8y_DistanceMeasurement": {"distance": {"value": 13.37, "unit": "mm"}}, "c8y_ThreePhase...	"c8yDistanceMeasurementdistancevalueunitmmc8yThreePhaseEnergyMeasurementA valueunitkWhA- valueuni...
Row5	"{"c8y_DistanceMeasurement": {"distance": {"value": 13.37, "unit": "mm"}}, "c8y_SinglePhase...	"c8yDistanceMeasurementdistancevalueunitmmc8ySinglePhaseEnergyMeasurementA valueunitkWhA- valueuni...
Row6	"{"c8y_VoltageMeasurement": {"voltage": {"value": 13.37, "unit": "V"}}, "c8y_LightMeasurem...	"c8yVoltageMeasurementvoltagevalueunitVc8yLightMeasurementе valueunitluxc8ySinglePhaseEnergyMeasure...
Row7	"{"c8y_DistanceMeasurement": {"distance": {"value": 13.37, "unit": "mm"}}, "c8y_LightMeasu...	"c8yDistanceMeasurementdistancevalueunitmmc8yLightMeasurementе valueunitluxc8yRelay relayState OPEN"
Row8	"{"c8y_HumidityMeasurement": {"h": {"value": 13.37, "unit": "%RH"}}, "c8y_LightMeasureme...	"c8yHumidityMeasurementh valueunitRHc8yLightMeasurementе valueunitluxc8yVoltageMeasurementvoltagev...
Row9	"{"c8y_ThreePhaseEnergyMeasurement": {"A+": {"value": 435, "unit": "kWh"}, "A-": {"value":...	"c8yThreePhaseEnergyMeasurementA valueunitkWhA- valueunitkWhP valueunitW P- valueunitW Avalueunitk...
Row10	"{"c8y_SinglePhaseEnergyMeasurement": {"A+": {"value": 123, "unit": "kWh"}, "A-": {"value":...	"c8ySinglePhaseEnergyMeasurementA valueunitkWhA- valueunitkWhP valueunitW P- valueunitWc8yDistance...
Row11	"{"c8y_RelayArray": ["OPEN", "CLOSED", "CLOSED", "OPEN"], "c8y_MotionMeasurement": {"...	"c8yRelayArray OPEN CLOSED CLOSED OPEN c8yMotionMeasurementmotionDetectedvalueunittype BOOLEAN...
Row12	"{"c8y_SinglePhaseEnergyMeasurement": {"A+": {"value": 123, "unit": "kWh"}, "A-": {"value":...	"c8ySinglePhaseEnergyMeasurementA valueunitkWhA- valueunitkWhP valueunitW P- valueunitWc8yThreePha...
Row13	"{"c8y_RelayArray": ["OPEN", "CLOSED", "CLOSED", "OPEN"], "c8y_VoltageMeasurement": {"...	"c8yRelayArray OPEN CLOSED CLOSED OPEN c8yVoltageMeasurementvoltagevalueunitVc8yLightMeasuremen...
Row14	"{"c8y_DistanceMeasurement": {"distance": {"value": 13.37, "unit": "mm"}}}"	"c8yDistanceMeasurementdistancevalueunitmm"
Row15	"{"c8y_VoltageMeasurement": {"voltage": {"value": 13.37, "unit": "V"}}, "c8y_CurrentMeasur...	"c8yVoltageMeasurementvoltagevalueunitVc8yCurrentMeasurementcurrentvalueunitAc8yRelayArray OPEN CL...
Row16	"{"c8y_Relay": {"relayState": "OPEN"}, "c8y_LightMeasurement": {"e": {"value": 8.36, "unit": ...	"c8yRelay relayState OPEN c8yLightMeasurementе valueunitluxc8yCurrentMeasurementcurrentvalueunitAc8y...
Row17	"{"c8y_SinglePhaseEnergyMeasurement": {"A+": {"value": 123, "unit": "kWh"}, "A-": {"value":...	"c8ySinglePhaseEnergyMeasurementA valueunitkWhA- valueunitkWhP valueunitW P- valueunitWc8yDistance...
Row18	"{"c8y_VoltageMeasurement": {"voltage": {"value": 13.37, "unit": "V"}}, "c8y_Relay": {"relay...	"c8yVoltageMeasurementvoltagevalueunitVc8yRelay relayState OPEN c8yDistanceMeasurementdistancevalue...
Row19	"{"c8y_LightMeasurement": {"e": {"value": 8.36, "unit": "lux"}}, "c8y_MotionMeasurement": {"...	"c8yLightMeasurementе valueunitluxc8yMotionMeasurementmotionDetectedvalueunittype BOOLEAN speedval...
Row20	"{"c8y_Relay": {"relayState": "OPEN"}, "c8y_MotionMeasurement": {"motionDetected": {"valu...	"c8yRelay relayState OPEN c8yMotionMeasurementmotionDetectedvalueunittype BOOLEAN speedvalueunitkm...
Row21	"{"c8y_Relay": {"relayState": "OPEN"}, "c8y_MotionMeasurement": {"motionDetected": {"valu...	"c8yRelay relayState OPEN c8yMotionMeasurementmotionDetectedvalueunittype BOOLEAN speedvalueunitkm...
Row22	"{"c8y_Relay": {"relayState": "OPEN"}, "c8y_DistanceMeasurement": {"distance": {"value": 1...	"c8yRelay relayState OPEN c8yDistanceMeasurementdistancevalueunitmm"
Row23	"{"c8y_LightMeasurement": {"e": {"value": 8.36, "unit": "lux"}}, "c8y_VoltageMeasurement": {...	"c8yLightMeasurementе valueunitluxc8yVoltageMeasurementvoltagevalueunitVc8yThreePhaseEnergyMeasure...
Row24	"{"c8y_ThreePhaseEnergyMeasurement": {"A+": {"value": 435, "unit": "kWh"}, "A-": {"value":...	"c8yThreePhaseEnergyMeasurementA valueunitkWhA- valueunitkWhP valueunitW P- valueunitW Avalueunitk...
Row25	"{"c8y_AccelerationMeasurement": {"acceleration": {"value": 8.36, "unit": "m/s2"}}}"	"c8yAccelerationMeasurementaccelerationvalueunitms2"
Row26	"{"c8y_Relay": {"relayState": "OPEN"}, "c8y_DistanceMeasurement": {"distance": {"value": 1...	"c8yRelay relayState OPEN c8yDistanceMeasurementdistancevalueunitmmc8yVoltageMeasurementvoltageval...
Row27	"{"c8y_HumidityMeasurement": {"h": {"value": 13.37, "unit": "%RH"}}, "c8y_MotionMeasurem...	"c8yHumidityMeasurementh valueunitRHc8yMotionMeasurementmotionDetectedvalueunittype BOOLEAN spee...
Row28	"{"c8y_ThreePhaseEnergyMeasurement": {"A+": {"value": 435, "unit": "kWh"}, "A-": {"value":...	"c8yThreePhaseEnergyMeasurementA valueunitkWhA- valueunitkWhP valueunitW P- valueunitW Avalueunitk...
Row29	"{"c8y_Relay": {"relayState": "OPEN"}}"	"c8yRelay relayState OPEN"
Row30	"{"c8y_ThreePhaseEnergyMeasurement": {"A+": {"value": 435, "unit": "kWh"}, "A-": {"value":...	"c8yThreePhaseEnergyMeasurementA valueunitkWhA- valueunitkWhP valueunitW P- valueunitW Avalueunitk...
Row31	"{"c8y_VoltageMeasurement": {"voltage": {"value": 13.37, "unit": "V"}}, "c8y_CurrentMeasur...	"c8yVoltageMeasurementvoltagevalueunitVc8yCurrentMeasurementcurrentvalueunitAc8yRelay relayState OP...
Row32	"{"c8y_CurrentMeasurement": {"current": {"value": 13.37, "unit": "A"}}, "c8y_HumidityMeasu...	"c8yCurrentMeasurementcurrentvalueunitAc8yHumidityMeasurementh valueunitRHc8yRelay relayState OPEN ...
Row33	"{"c8y_AccelerationMeasurement": {"acceleration": {"value": 8.36, "unit": "m/s2"}}, "c8y_Ligh...	"c8yAccelerationMeasurementaccelerationvalueunitms2c8yLightMeasurementе valueunitluxc8yMotionMeasure...
Row34	"{"c8y_SinglePhaseEnergyMeasurement": {"A+": {"value": 123, "unit": "kWh"}, "A-": {"value":...	"c8ySinglePhaseEnergyMeasurementA valueunitkWhA- valueunitkWhP valueunitW P- valueunitWc8yRelay rela...
Row35	"{"c8y_HumidityMeasurement": {"h": {"value": 13.37, "unit": "%RH"}}, "c8y_AccelerationMea...	"c8yHumidityMeasurementh valueunitRHc8yAccelerationMeasurementaccelerationvalueunitms2c8yRelayArray...
Row36	"{"c8y_RelayArray": ["OPEN", "CLOSED", "CLOSED", "OPEN"], "c8y_ThreePhaseEnergyMeasu...	"c8yRelayArray OPEN CLOSED CLOSED OPEN c8yThreePhaseEnergyMeasurementA valueunitkWhA- valueunit...
Row37	"{"c8y_AccelerationMeasurement": {"acceleration": {"value": 8.36, "unit": "m/s2"}}}"	"c8yAccelerationMeasurementaccelerationvalueunitms2"

Figure 17: Number Filter Output (Ashari, 2021)

Using the Bag-of-Words (BoW) creator node, the next phase of the ML pipeline is to build a term column consisting of each word in the document rows.

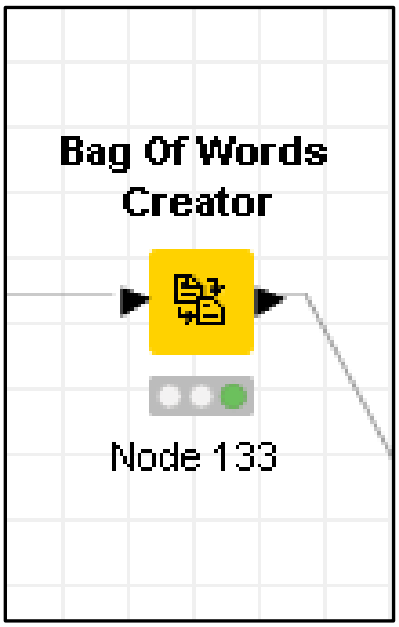


Figure 18: Bags-of-Words Creator Node (Ashari, 2021)

Documents output table - 3:133 - Bag Of Words Creator

File Edit Hilite Navigation View

Table "default" - Rows: 30442 | Spec - Columns: 2 | Properties | Flow Variables

Row ID	Preprocessed Document	Term
Row14031	"c8yRelay relayState OPEN c8yLightMeasuremente valueunitluxc8ySinglePhaseEn...	Wh[]
Row14032	"c8yRelay relayState OPEN c8yLightMeasuremente valueunitluxc8ySinglePhaseEn...	A-[]
Row14033	"c8yRelay relayState OPEN c8yLightMeasuremente valueunitluxc8ySinglePhaseEn...	P[]
Row14034	"c8yRelay relayState OPEN c8yLightMeasuremente valueunitluxc8ySinglePhaseEn...	W[]
Row14035	"c8yRelay relayState OPEN c8yLightMeasuremente valueunitluxc8ySinglePhaseEn...	P-[]
Row14036	"c8yRelay relayState OPEN c8yLightMeasuremente valueunitluxc8ySinglePhaseEn...	c8yPosition[]
Row14037	"c8yRelay relayState OPEN c8yLightMeasuremente valueunitluxc8ySinglePhaseEn...	alt[]
Row14038	"c8yRelay relayState OPEN c8yLightMeasuremente valueunitluxc8ySinglePhaseEn...	lng[]
Row14039	"c8yRelay relayState OPEN c8yLightMeasuremente valueunitluxc8ySinglePhaseEn...	lat[]
Row14040	"c8yRelay relayState OPEN c8yLightMeasuremente valueunitluxc8ySinglePhaseEn...	trackingProtocol[]
Row14041	"c8yRelay relayState OPEN c8yLightMeasuremente valueunitluxc8ySinglePhaseEn...	TELIC[]
Row14042	"c8yRelay relayState OPEN c8yLightMeasuremente valueunitluxc8ySinglePhaseEn...	reportReason[]
Row14043	"c8yRelay relayState OPEN c8yLightMeasuremente valueunitluxc8ySinglePhaseEn...	Time[]
Row14044	"c8yRelay relayState OPEN c8yLightMeasuremente valueunitluxc8ySinglePhaseEn...	Event[]
Row14045	"c8yRelay relayState OPEN c8yLightMeasuremente valueunitluxc8ySinglePhaseEn...	c8yDistanceMeasurement[]
Row14046	"c8yRelay relayState OPEN c8yLightMeasuremente valueunitluxc8ySinglePhaseEn...	distance[]
Row14047	"c8yRelay relayState OPEN c8yLightMeasuremente valueunitluxc8ySinglePhaseEn...	mm[]
Row14048	"c8yPositionaltlng lat trackingProtocol TELIC reportReasonTime Eventc8yDistance...	c8yPosition[]
Row14049	"c8yPositionaltlng lat trackingProtocol TELIC reportReasonTime Eventc8yDistance...	alt[]
Row14050	"c8yPositionaltlng lat trackingProtocol TELIC reportReasonTime Eventc8yDistance...	lng[]
Row14051	"c8yPositionaltlng lat trackingProtocol TELIC reportReasonTime Eventc8yDistance...	lat[]
Row14052	"c8yPositionaltlng lat trackingProtocol TELIC reportReasonTime Eventc8yDistance...	trackingProtocol[]
Row14053	"c8yPositionaltlng lat trackingProtocol TELIC reportReasonTime Eventc8yDistance...	TELIC[]
Row14054	"c8yPositionaltlng lat trackingProtocol TELIC reportReasonTime Eventc8yDistance...	reportReason[]
Row14055	"c8yPositionaltlng lat trackingProtocol TELIC reportReasonTime Eventc8yDistance...	Time[]

Figure 19: Bags-of-Words (BoW) Creator Output (Ashari, 2021)

Next, the Term Frequency (TF) node computes the term frequency of each term according to each document and summarises the result in the new column called TF abs (KNIME, 2021).

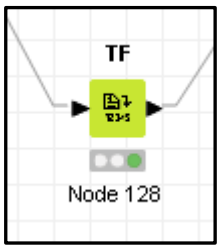


Figure 20: Term Frequency TF (Ashari, 2021)

Terms and documents output table - 3:128 - TF

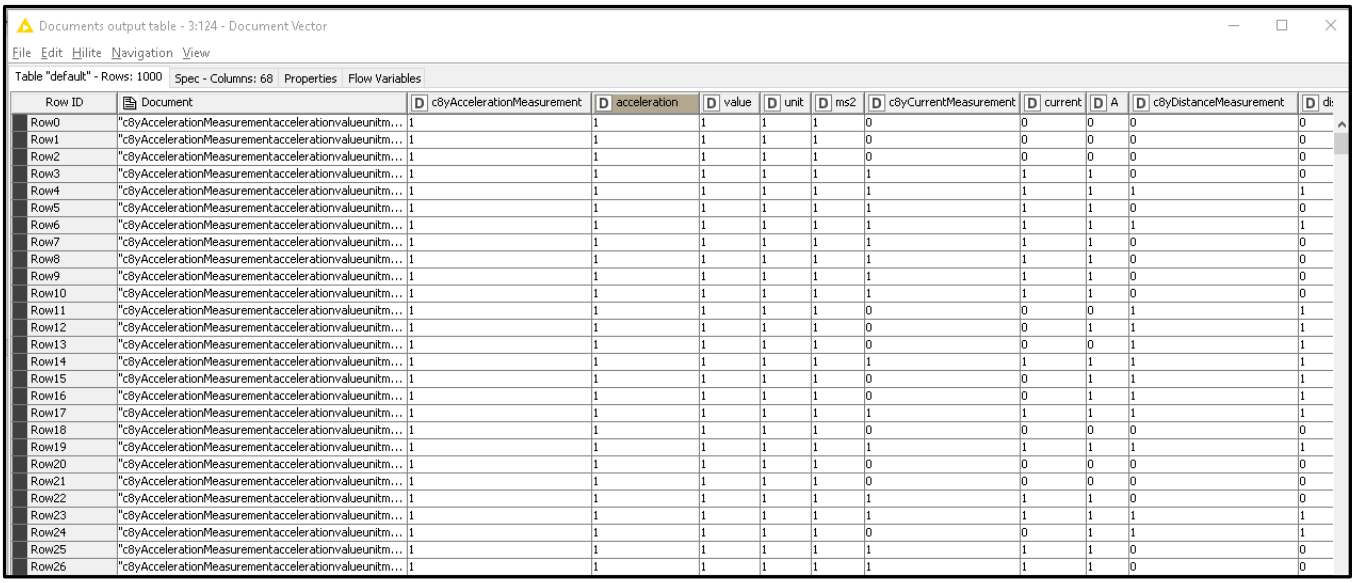
File Edit Hilite Navigation View

Table "default" - Rows: 30442 | Spec - Columns: 3 | Properties | Flow Variables

Row ID	Preprocessed Document	Term	TF abs
Row0	"c8yMotionMeasurementmotionDetectedvalueunittype BOOLEAN speedvalueunitkmh"	c8yMotionMeasureme...	2
Row1	"c8yMotionMeasurementmotionDetectedvalueunittype BOOLEAN speedvalueunitkmh"	motionDetected[]	2
Row2	"c8yMotionMeasurementmotionDetectedvalueunittype BOOLEAN speedvalueunitkmh"	value[]	4
Row3	"c8yMotionMeasurementmotionDetectedvalueunittype BOOLEAN speedvalueunitkmh"	unit[]	4
Row4	"c8yMotionMeasurementmotionDetectedvalueunittype BOOLEAN speedvalueunitkmh"	type[]	2
Row5	"c8yMotionMeasurementmotionDetectedvalueunittype BOOLEAN speedvalueunitkmh"	BOOLEAN[]	2
Row6	"c8yMotionMeasurementmotionDetectedvalueunittype BOOLEAN speedvalueunitkmh"	speed[]	2
Row7	"c8yMotionMeasurementmotionDetectedvalueunittype BOOLEAN speedvalueunitkmh"	kmh[]	2
Row8	"c8yAccelerationMeasurementaccelerationvalueunitms2c8yMotionMeasurementmotionDetectedvalueunittype BOOLEAN speedvalueunitkmhc8yVoltageMeasurementvoltagevalueunitVc8yRelay relayState OPEN"	c8yAccelerationMeasu...	2
Row9	"c8yAccelerationMeasurementaccelerationvalueunitms2c8yMotionMeasurementmotionDetectedvalueunittype BOOLEAN speedvalueunitkmhc8yVoltageMeasurementvoltagevalueunitVc8yRelay relayState OPEN"	acceleration[]	2
Row10	"c8yAccelerationMeasurementaccelerationvalueunitms2c8yMotionMeasurementmotionDetectedvalueunittype BOOLEAN speedvalueunitkmhc8yVoltageMeasurementvoltagevalueunitVc8yRelay relayState OPEN"	value[]	8
Row11	"c8yAccelerationMeasurementaccelerationvalueunitms2c8yMotionMeasurementmotionDetectedvalueunittype BOOLEAN speedvalueunitkmhc8yVoltageMeasurementvoltagevalueunitVc8yRelay relayState OPEN"	unit[]	8
Row12	"c8yAccelerationMeasurementaccelerationvalueunitms2c8yMotionMeasurementmotionDetectedvalueunittype BOOLEAN speedvalueunitkmhc8yVoltageMeasurementvoltagevalueunitVc8yRelay relayState OPEN"	ms2[]	2
Row13	"c8yAccelerationMeasurementaccelerationvalueunitms2c8yMotionMeasurementmotionDetectedvalueunittype BOOLEAN speedvalueunitkmhc8yVoltageMeasurementvoltagevalueunitVc8yRelay relayState OPEN"	c8yMotionMeasureme...	2
Row14	"c8yAccelerationMeasurementaccelerationvalueunitms2c8yMotionMeasurementmotionDetectedvalueunittype BOOLEAN speedvalueunitkmhc8yVoltageMeasurementvoltagevalueunitVc8yRelay relayState OPEN"	motionDetected[]	2
Row15	"c8yAccelerationMeasurementaccelerationvalueunitms2c8yMotionMeasurementmotionDetectedvalueunittype BOOLEAN speedvalueunitkmhc8yVoltageMeasurementvoltagevalueunitVc8yRelay relayState OPEN"	type[]	2
Row16	"c8yAccelerationMeasurementaccelerationvalueunitms2c8yMotionMeasurementmotionDetectedvalueunittype BOOLEAN speedvalueunitkmhc8yVoltageMeasurementvoltagevalueunitVc8yRelay relayState OPEN"	BOOLEAN[]	2
Row17	"c8yAccelerationMeasurementaccelerationvalueunitms2c8yMotionMeasurementmotionDetectedvalueunittype BOOLEAN speedvalueunitkmhc8yVoltageMeasurementvoltagevalueunitVc8yRelay relayState OPEN"	speed[]	2
Row18	"c8yAccelerationMeasurementaccelerationvalueunitms2c8yMotionMeasurementmotionDetectedvalueunittype BOOLEAN speedvalueunitkmhc8yVoltageMeasurementvoltagevalueunitVc8yRelay relayState OPEN"	kmh[]	2
Row19	"c8yAccelerationMeasurementaccelerationvalueunitms2c8yMotionMeasurementmotionDetectedvalueunittype BOOLEAN speedvalueunitkmhc8yVoltageMeasurementvoltagevalueunitVc8yRelay relayState OPEN"	c8yVoltageMeasureme...	2
Row20	"c8yAccelerationMeasurementaccelerationvalueunitms2c8yMotionMeasurementmotionDetectedvalueunittype BOOLEAN speedvalueunitkmhc8yVoltageMeasurementvoltagevalueunitVc8yRelay relayState OPEN"	voltage[]	2
Row21	"c8yAccelerationMeasurementaccelerationvalueunitms2c8yMotionMeasurementmotionDetectedvalueunittype BOOLEAN speedvalueunitkmhc8yVoltageMeasurementvoltagevalueunitVc8yRelay relayState OPEN"	V[]	2
Row22	"c8yAccelerationMeasurementaccelerationvalueunitms2c8yMotionMeasurementmotionDetectedvalueunittype BOOLEAN speedvalueunitkmhc8yVoltageMeasurementvoltagevalueunitVc8yRelay relayState OPEN"	c8yRelay[]	2
Row23	"c8yAccelerationMeasurementaccelerationvalueunitms2c8yMotionMeasurementmotionDetectedvalueunittype BOOLEAN speedvalueunitkmhc8yVoltageMeasurementvoltagevalueunitVc8yRelay relayState OPEN"	relayState[]	2
Row24	"c8yAccelerationMeasurementaccelerationvalueunitms2c8yMotionMeasurementmotionDetectedvalueunittype BOOLEAN speedvalueunitkmhc8yVoltageMeasurementvoltagevalueunitVc8yRelay relayState OPEN"	OPEN[]	2

Figure 21: TF Output (Ashari, 2021)

The document vector node registers each word's simple presence, summarising the full message as a vector.



Documents output table - 3:124 - Document Vector

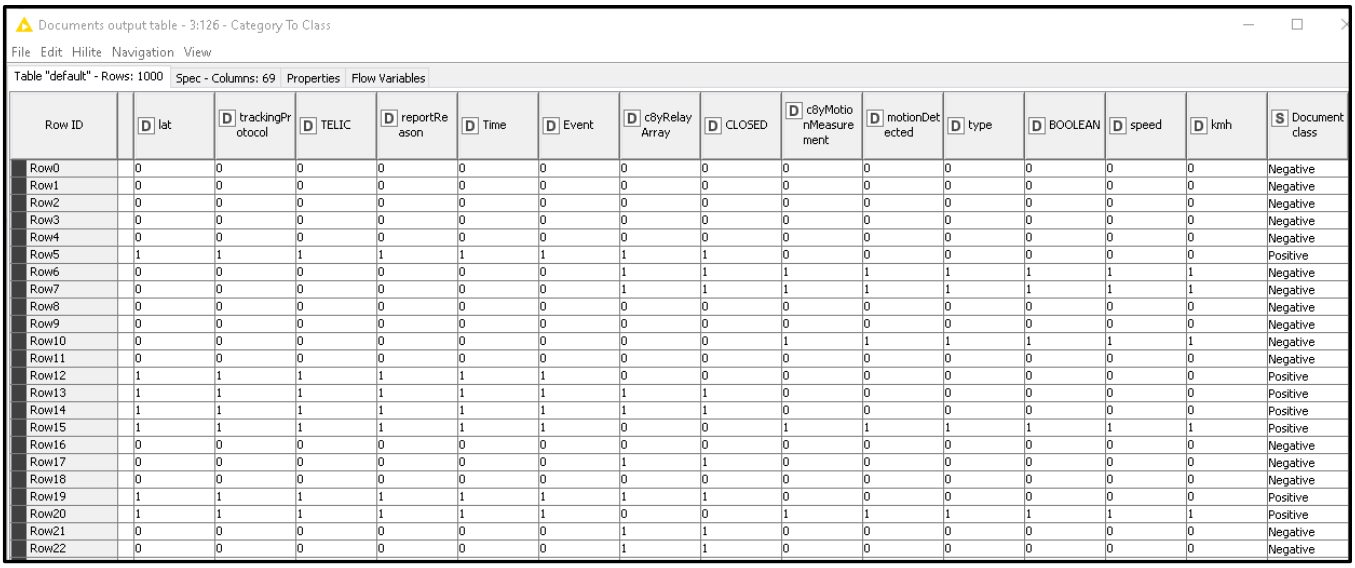
File Edit Hilite Navigation View

Table "default" - Rows: 1000 | Spec - Columns: 68 | Properties | Flow Variables

Row ID	Document	c8yAccelerationMeasurement	acceleration	value	unit	ms2	c8yCurrentMeasurement	current	A	c8yDistanceMeasurement	di
Row0	"c8yAccelerationMeasurementaccelerationvalueunitm...	1	1	1	1	1	0	0	0	0	0
Row1	"c8yAccelerationMeasurementaccelerationvalueunitm...	1	1	1	1	1	0	0	0	0	0
Row2	"c8yAccelerationMeasurementaccelerationvalueunitm...	1	1	1	1	1	0	0	0	0	0
Row3	"c8yAccelerationMeasurementaccelerationvalueunitm...	1	1	1	1	1	1	1	1	0	0
Row4	"c8yAccelerationMeasurementaccelerationvalueunitm...	1	1	1	1	1	1	1	1	1	1
Row5	"c8yAccelerationMeasurementaccelerationvalueunitm...	1	1	1	1	1	1	1	1	0	0
Row6	"c8yAccelerationMeasurementaccelerationvalueunitm...	1	1	1	1	1	1	1	1	1	1
Row7	"c8yAccelerationMeasurementaccelerationvalueunitm...	1	1	1	1	1	1	1	1	0	0
Row8	"c8yAccelerationMeasurementaccelerationvalueunitm...	1	1	1	1	1	1	1	1	0	0
Row9	"c8yAccelerationMeasurementaccelerationvalueunitm...	1	1	1	1	1	1	1	1	0	0
Row10	"c8yAccelerationMeasurementaccelerationvalueunitm...	1	1	1	1	1	1	1	1	0	0
Row11	"c8yAccelerationMeasurementaccelerationvalueunitm...	1	1	1	1	1	0	0	0	1	1
Row12	"c8yAccelerationMeasurementaccelerationvalueunitm...	1	1	1	1	1	0	0	1	1	1
Row13	"c8yAccelerationMeasurementaccelerationvalueunitm...	1	1	1	1	1	0	0	0	1	1
Row14	"c8yAccelerationMeasurementaccelerationvalueunitm...	1	1	1	1	1	1	1	1	1	1
Row15	"c8yAccelerationMeasurementaccelerationvalueunitm...	1	1	1	1	1	0	0	1	1	1
Row16	"c8yAccelerationMeasurementaccelerationvalueunitm...	1	1	1	1	1	0	0	1	1	1
Row17	"c8yAccelerationMeasurementaccelerationvalueunitm...	1	1	1	1	1	1	1	1	1	1
Row18	"c8yAccelerationMeasurementaccelerationvalueunitm...	1	1	1	1	1	0	0	0	0	0
Row19	"c8yAccelerationMeasurementaccelerationvalueunitm...	1	1	1	1	1	1	1	1	1	1
Row20	"c8yAccelerationMeasurementaccelerationvalueunitm...	1	1	1	1	1	0	0	0	0	0
Row21	"c8yAccelerationMeasurementaccelerationvalueunitm...	1	1	1	1	1	0	0	0	0	0
Row22	"c8yAccelerationMeasurementaccelerationvalueunitm...	1	1	1	1	1	1	1	1	0	0
Row23	"c8yAccelerationMeasurementaccelerationvalueunitm...	1	1	1	1	1	1	1	1	1	1
Row24	"c8yAccelerationMeasurementaccelerationvalueunitm...	1	1	1	1	1	0	0	1	1	1
Row25	"c8yAccelerationMeasurementaccelerationvalueunitm...	1	1	1	1	1	1	1	1	0	0
Row26	"c8yAccelerationMeasurementaccelerationvalueunitm...	1	1	1	1	1	1	1	1	0	0

Figure 22: Document Vector Output (Ashari, 2021)

The category-to-class node mapped the rows into the document class based on the earlier label assigned to the input files (negative or positive). Basically, in the document that has cy8_Position information, this node adds a positive label, and if there is no cy8_Position information in it, this node adds a negative label. With this result, we are ready to train our ML model.



Documents output table - 3:126 - Category To Class

File Edit Hilite Navigation View

Table "default" - Rows: 1000 | Spec - Columns: 69 | Properties | Flow Variables

Row ID	lat	trackingProtocol	TELIC	reportReason	Time	Event	c8yRelayArray	CLOSED	c8yMotionMeasurement	motionDetected	type	BOOLEAN	speed	kmh	Document class
Row0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Negative
Row1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Negative
Row2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Negative
Row3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Negative
Row4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Negative
Row5	1	1	1	1	1	1	1	1	0	0	0	0	0	0	Positive
Row6	0	0	0	0	0	0	1	1	1	1	1	1	1	1	Negative
Row7	0	0	0	0	0	0	1	1	1	1	1	1	1	1	Negative
Row8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Negative
Row9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Negative
Row10	0	0	0	0	0	0	0	0	1	1	1	1	1	1	Negative
Row11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Negative
Row12	1	1	1	1	1	1	0	0	0	0	0	0	0	0	Positive
Row13	1	1	1	1	1	1	1	1	0	0	0	0	0	0	Positive
Row14	1	1	1	1	1	1	1	1	0	0	0	0	0	0	Positive
Row15	1	1	1	1	1	1	0	0	1	1	1	1	1	1	Positive
Row16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Negative
Row17	0	0	0	0	0	0	1	1	0	0	0	0	0	0	Negative
Row18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Negative
Row19	1	1	1	1	1	1	1	1	0	0	0	0	0	0	Positive
Row20	1	1	1	1	1	1	0	0	1	1	1	1	1	1	Positive
Row21	0	0	0	0	0	0	1	1	0	0	0	0	0	0	Negative
Row22	0	0	0	0	0	0	1	1	0	0	0	0	0	0	Negative

Figure 23: Category-to-Class output (Ashari, 2021)

The next step is to train the model using supervised learning, a binary classifier, in this case, the Decision Tree Learner. The study used the partitioner node to partition the training and test dataset data in a 70:30 ratio. Once the model learns this classifier, the Decision Tree Predictor can be fed with JSON messages to determine their label (**Document Class = Positive** if it contains PII, or **Document Class = Negative** if PII does not exist).

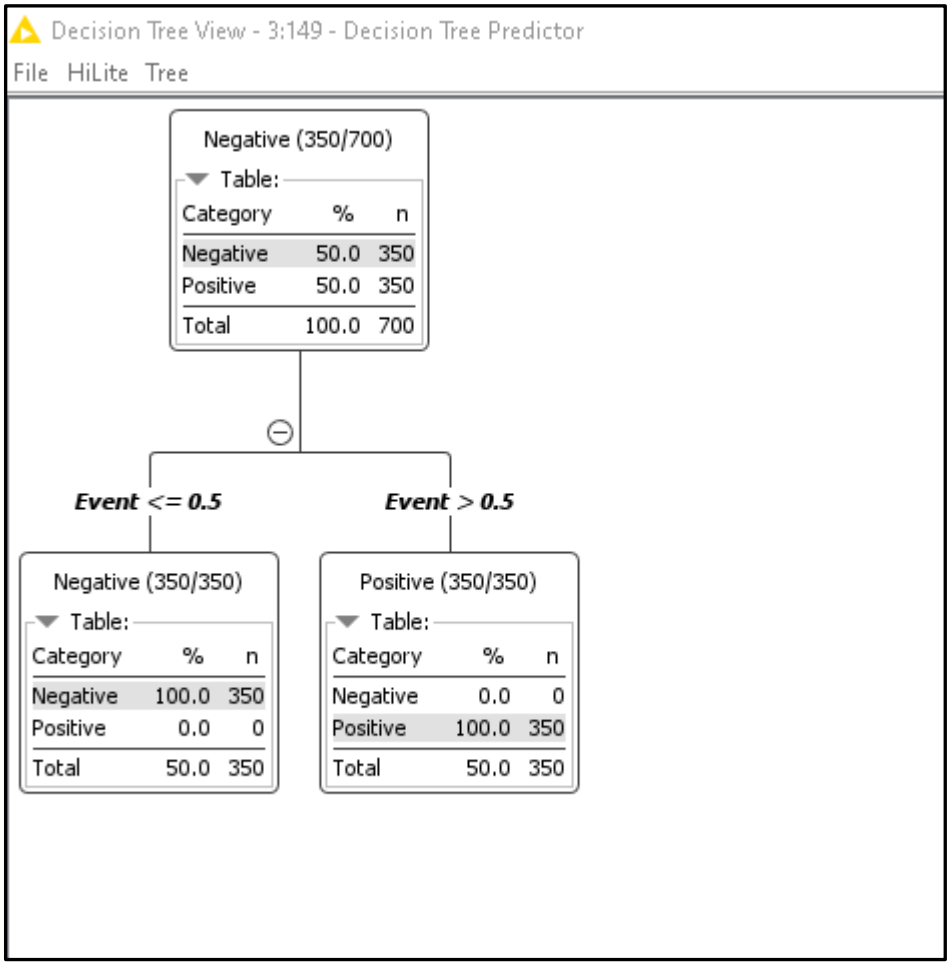


Figure 24: Decision Tree Predictor (Ashari, 2021)

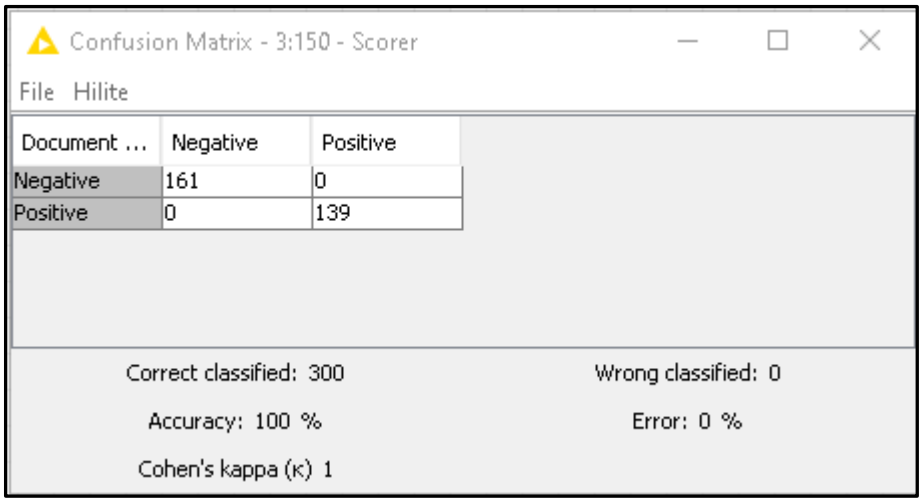
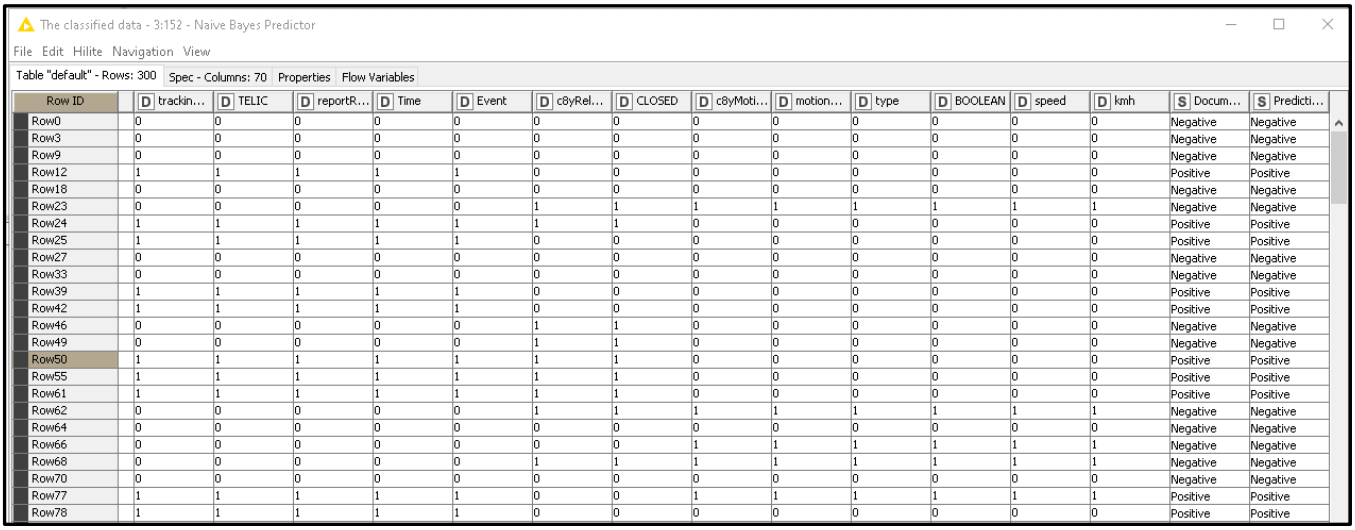


Figure 25: Scorer Output (Ashari, 2021)

Figure 24 shows the Decision Tree Predictor node's negative and positive document predictions. From figure 25, the scorer node shows a 100% accuracy confusion matrix.

The Decision Tree classifier is our baseline for this ML model. The second classifier, which is Naïve Bayes, is used to check whether the outcome from the Decision Tree classifier is consistent. Figure 26 shows the output from the Naïve Bayes predictor.



Row ID	trackin...	TELIC	reportR...	Time	Event	c8yRel...	CLOSED	c8yMoti...	motion...	type	BOOLEAN	speed	kmh	Docum...	Predicti...
Row0	0	0	0	0	0	0	0	0	0	0	0	0	0	Negative	Negative
Row3	0	0	0	0	0	0	0	0	0	0	0	0	0	Negative	Negative
Row9	0	0	0	0	0	0	0	0	0	0	0	0	0	Negative	Negative
Row12	1	1	1	1	1	0	0	0	0	0	0	0	0	Positive	Positive
Row18	0	0	0	0	0	0	0	0	0	0	0	0	0	Negative	Negative
Row23	0	0	0	0	0	1	1	1	1	1	1	1	1	Negative	Negative
Row24	1	1	1	1	1	1	1	0	0	0	0	0	0	Positive	Positive
Row25	1	1	1	1	1	0	0	0	0	0	0	0	0	Positive	Positive
Row27	0	0	0	0	0	0	0	0	0	0	0	0	0	Negative	Negative
Row33	0	0	0	0	0	0	0	0	0	0	0	0	0	Negative	Negative
Row39	1	1	1	1	1	0	0	0	0	0	0	0	0	Positive	Positive
Row42	1	1	1	1	1	0	0	0	0	0	0	0	0	Positive	Positive
Row46	0	0	0	0	0	1	1	0	0	0	0	0	0	Negative	Negative
Row49	0	0	0	0	0	1	1	0	0	0	0	0	0	Negative	Negative
Row50	1	1	1	1	1	1	1	0	0	0	0	0	0	Positive	Positive
Row55	1	1	1	1	1	1	1	0	0	0	0	0	0	Positive	Positive
Row61	1	1	1	1	1	1	1	0	0	0	0	0	0	Positive	Positive
Row62	0	0	0	0	0	1	1	1	1	1	1	1	1	Negative	Negative
Row64	0	0	0	0	0	0	0	0	0	0	0	0	0	Negative	Negative
Row66	0	0	0	0	0	0	0	1	1	1	1	1	1	Negative	Negative
Row68	0	0	0	0	0	1	1	1	1	1	1	1	1	Negative	Negative
Row70	0	0	0	0	0	0	0	0	0	0	0	0	0	Negative	Negative
Row77	1	1	1	1	1	0	0	1	1	1	1	1	1	Positive	Positive
Row78	1	1	1	1	1	0	0	0	0	0	0	0	0	Positive	Positive

Figure 26: Naive Bayes Predictor Output (Ashari, 2021)

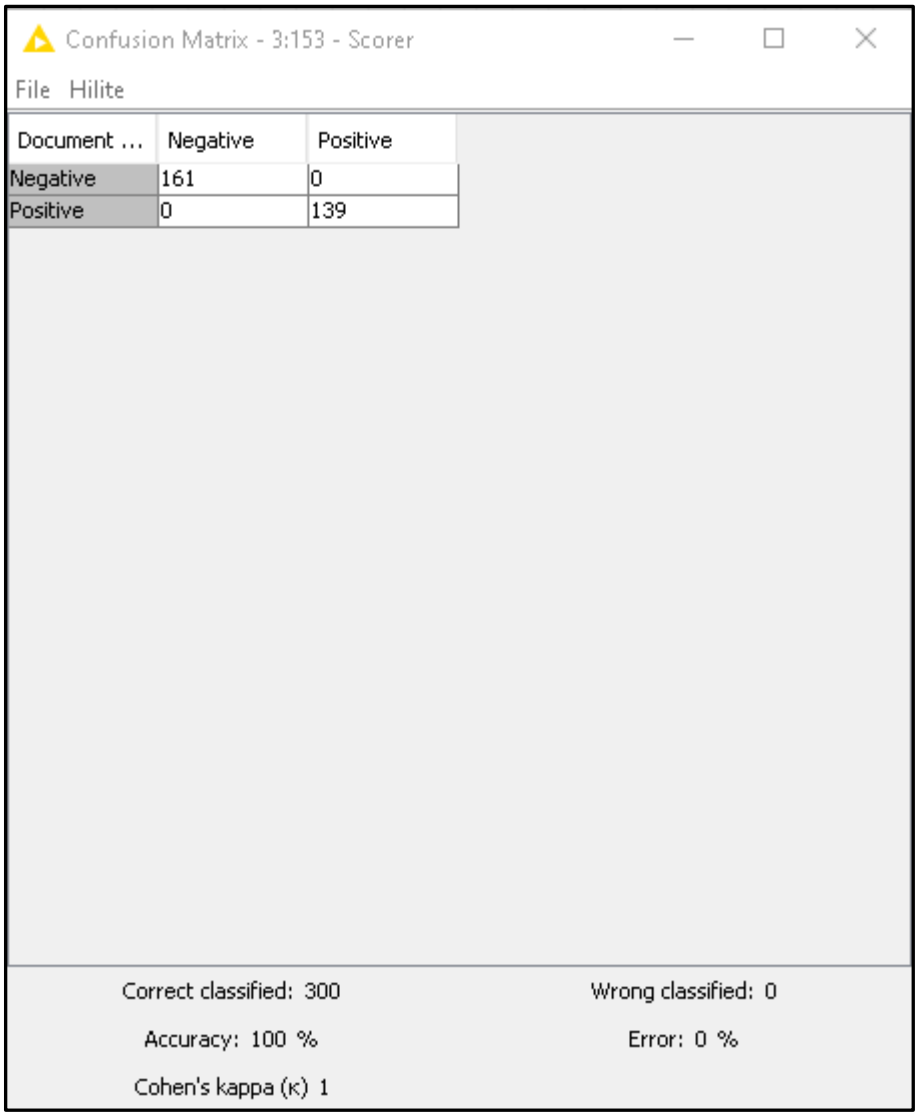


Figure 27: Scorer Output for Naive Bayes Predictor (Ashari, 2021)

The result from the Naïve Bayes classifier confirms the result of the Decision Tree classifier.

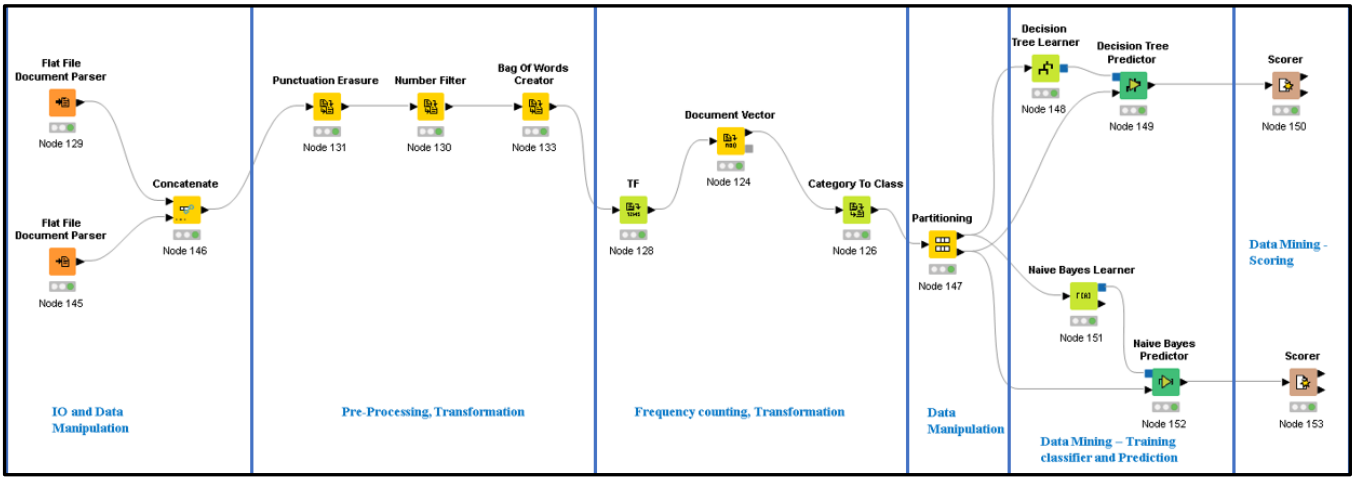


Figure 28: End-to-End ML Pipeline (Ashari, 2021)

3.6 RB Modelling and Classification

This section analysed the IoT payload data format from the RB modelling and classification from the data dump side. Based on the Cumolocity documentation, the data dump will be in a stream of rows in a comma-separated-values (CSV) format.

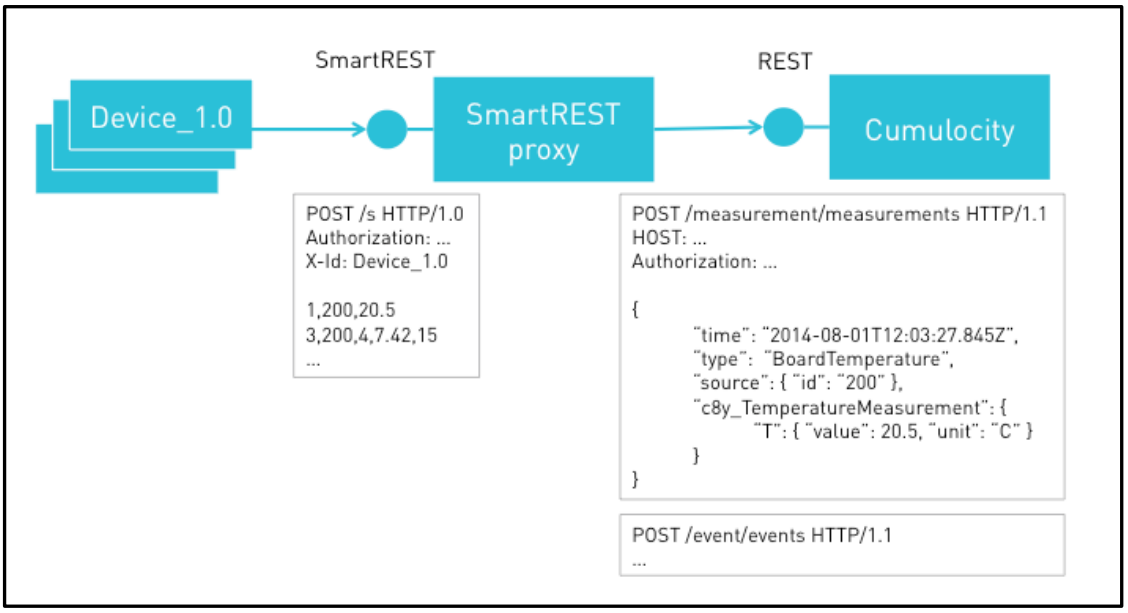


Figure 29: Example of Temperature Sensor Data Dump (Cumulocity, 2021)

Since we do not have access to the raw data dump sample from an IoT device for this study, we will use instead already formatted data similar to the one that originated from a SmartREST proxy. This rule-based condition will classify PII data if fragment name cy8_Position exists in the IoT data payload. In figure 30, a program executing a rule-based process will detect in the incoming JSON formatted data PII keys that are matched against a list of keys, in our case, the cy8_Position.

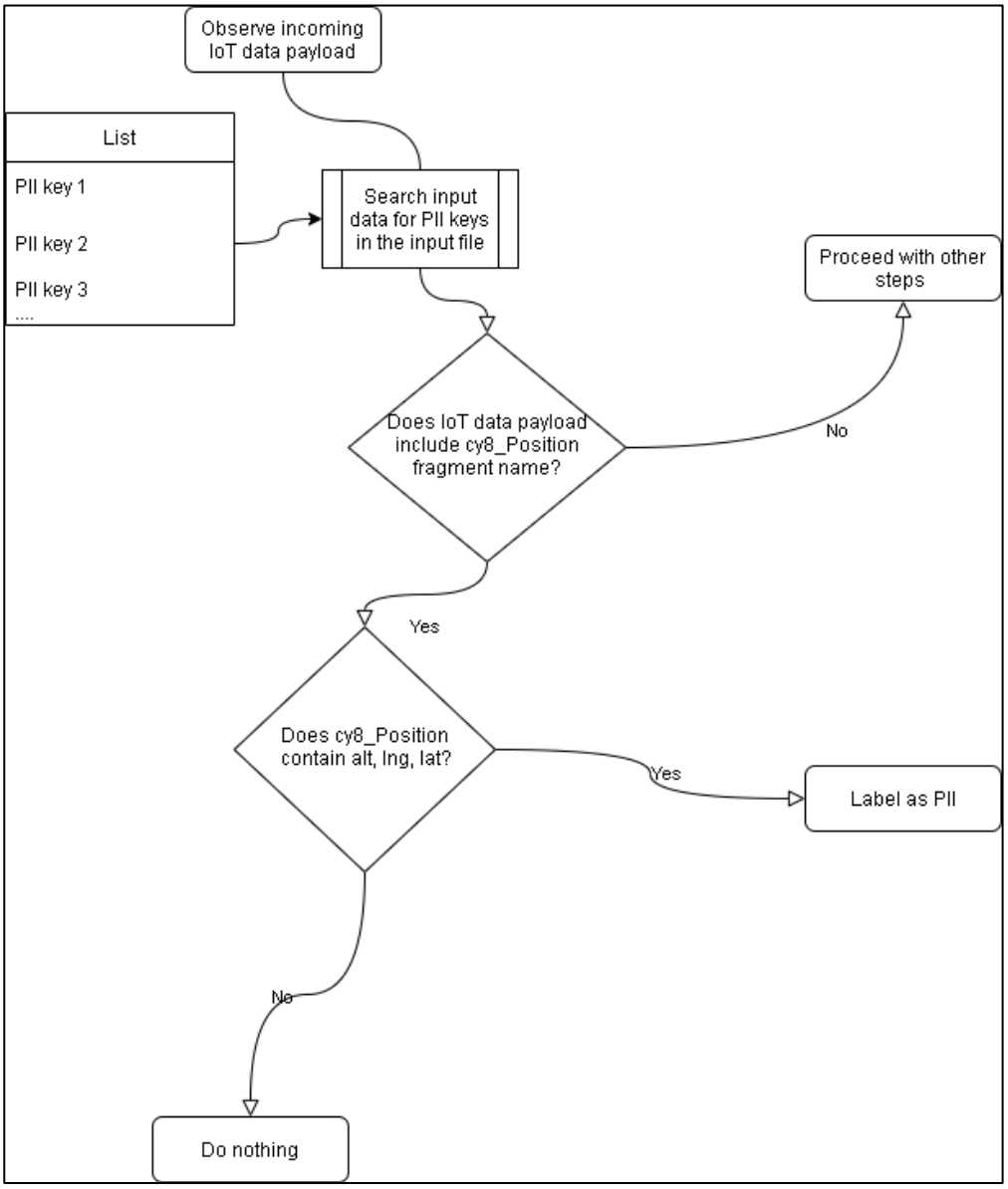


Figure 30: IoT Data Stream to SmartREST API (Ashari, 2021)

Next, this study utilised a regular expression tool to label the desired location data from the JSON generated input files containing many rows of fragment names. Figure 31 demonstrated that we could capture a group having the rule-based condition of the location data referred by PII. The regular expression, also known as regex, can capture any cy8_Postion PII, regardless of specific values contained in the altitude, longitude, and latitude. The example below shows a manual regex process. To automate the process, in the incoming IoT payload, the regex script can process data before SmartREST API takes place to capture and tag the incoming targeted PII data.

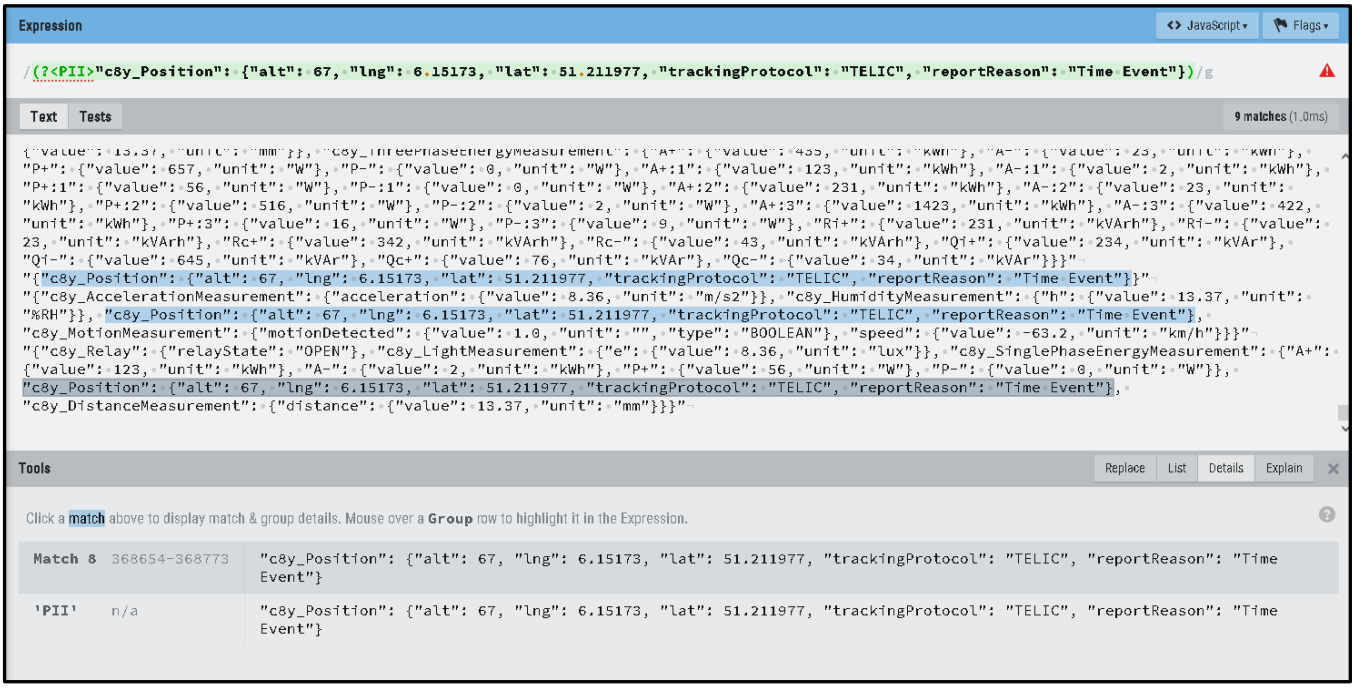


Figure 31: Rule-Based Output from Regular Expression (Ashari, 2021)

3.7 Observation on ML and RB Approaches

This sub-section summarises the limitations of each of the two methods in detecting private data from the IoT data payload.

The location in the data flow between IoT devices messages and the Cumulocity engine to fulfil their purpose is a key property of the RB and ML approaches.

The RB method of private data detection could take place before the IoT data payload reaches the SmartREST API as well as after it leaves the SmartREST API. The example in figure 31 above occurred after SmartREST API converted the data into the JSON file format. Knowing the exact format of the PII data structure, the RB method could detect which data belongs to PII and tag it accordingly, even in the IoT payload CSV format. By contrast, the ML method requires labelled data at the beginning of processing; this can happen only after SmartREST translates the message of the data payload into a JSON file format. However, after that, fully-trained ML processing no longer requires labelled input data.

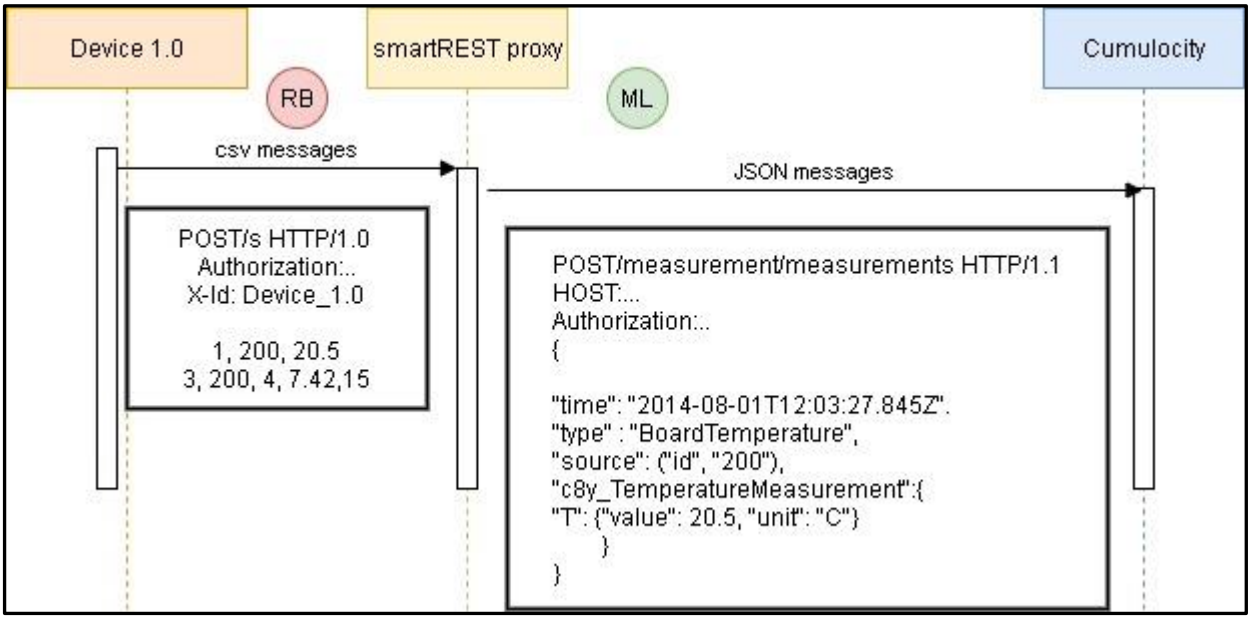


Figure 32: RB vs ML Method in the Incoming Data Flow to CoT (Ashari, 2021)

The advantage of the RB method lies in the fact that it could take place in the early stage. The earlier private data detection occurs, the better.

The second point is that the RB method requires an exact PII data format. The RB conditions need an accurate formula for correct data filtering. When changes happen, such as advancing new IoT technology or a new format in the IoT data payload, the RB algorithm needs to be updated, making it more difficult to maintain. Any small adjustments in the input file that is not matched with an updated RB filter may result in false-negative for the RB method. The ML approach lets the machine re-learn the model that suits the new PII-related data and improves their learning results. The ML method does not need an explicit definition of the PIIs but only the correct labelling to train the model.

The advantage of learning via ML is that the algorithm can improve by training the model with lots of data. The ML method is more robust due to the file preprocessing and bags-of-words nodes in the preprocessing and transformation phase. According to Azure IoT hub documentation, about 256-KB sensor data batches are sent by IoT devices every five minutes.[6] The first challenge is to create the right model for ML and continue to train the model with more data so that the algorithm will improve. As for the RB method, the IF-THEN conditions need to be up to date as the new data with the new PII format comes into the data stream.

Another observation point from this study's analysis is the amount of data required for each method to function as expected. The RB method does not require a large amount of data to validate the conditions specified in the rules. On the contrary, the ML method must follow iterations of learning and training to get the algorithm adjusted to the chosen ML model and improve the outcomes.

Finally, it is important to note system performance during ML or RB detecting the PII data. The RB method gave an almost instant output for the specified regex conditions. On the other hand, the ML approach is more resource

[6] https://docs.microsoft.com/en-us/azure/iot-hub/iot-hub-devguide-d2c-guidance

intensive because files needed to be converted into vectors to feed the classifiers. Note that we only sampled the study of the PII detection using one PII characteristic, which is the location data.

3.8 Chapter Summary

This chapter presented the analysis and the design of the two methods for the identified problems and described the ML and RB approaches and observations.

Chapter 4. Results and Evaluation

4.1 Chapter Introduction

This chapter aims to evaluate the outcome of ML and RB proposed workflows to detect PII data. We designed the assessment of this research's analysis and design outcome by interviewing the experts from DT personnel. Since this research topic addresses an existing gap in the IoT privacy offering, it makes sense to invite the company personnel to give their opinions and feedback on the achieved results. The interview took place in a teleconferencing session, and we received written interview feedback. There were altogether eight interview questions. The first question relates to the research question of whether it fits in the DT environment. The second and third questions explained the ML and the RB approaches and requested feedback regarding opportunities for and obstacles to implementing this method in DT. The subsequent questions looked into whether the research would be interesting and relevant for private data security, the privacy events, and if the solution could assist in the private data breaches. For reference, the interview questions are available in the appendix section.

For validation on the ML and RB approaches, we interviewed a senior Artificial Intelligence (AI) Manager with over twenty years of experience in enterprise architecture and product development. In this session, we demonstrated the ML pipeline in detail and the RB method outcome to the AI manager.

The second interview focused on the privacy aspect of this research. We have interviewed a senior Privacy and Security Manager with over five years of working experience in the Privacy department. In this session, we went through the entire thesis document, concentrating on the research questions, the motivations, how we designed the ML and RB approach, and the outcomes from these studies.

4.2 Evaluation of the ML Test Results

This study used two classifiers with different ML algorithms to train the model and predict the input data. Decision Tree is a fast tree-based model with automatic feature selections (Verwer, 2021). The Naïve Bayes classifier is a linear classifier with an easy-to-compute probabilistic model known to be highly accurate when applied to big data. Implementing these two ML algorithms to the IoT payload SmartREST training data produces similar results.

Both classifier scorers gave a result with 100% accuracy, showing that the training data was overfitting. Overfitting is when the training data fit the desired outcome too precisely, leading to poor prediction for new data (Verwer, 2021). To overcome this overfitting scenario, we should train the model with the actual or close to the real IoT data payload from the CoT. The result can lead to one of two conclusions: the ML approach does not apply to the chosen scenario, or it shows that we need to train the model with real data.

In an interview regarding ML pipeline and research questions, the DT AI expert agreed that both the ML classifier's design and outcome are valid with the available training data generated by Python scripts. DT can

view the result as their first baseline. The algorithm can be further enhanced with different ML models and implemented in a more enterprise-like environment with ML automation and operations teams.

It is useful to add the computational cost calculation to compare two classifiers in the ML pipeline. However, since we did not use the real test data, measuring execution time from randomly generated data is irrelevant. It is important to compute the processing time when the ML algorithm runs in an enterprise-like environment.

The DT AI expert also commented on the possible obstacle to implementing the ML approach to detect private data leaks in DT CoT. According to him, implementing the ML approach to protect PII in one organization (such as CoT) without considering the horizontal and vertical relationships between organizations in the company will result in confusion and inefficiencies, i.e., the silo effect.

We also interviewed the DT Privacy expert. He believed that ML can detect anomalies in the IoT payload. He sees the training and the ML model maintenance as the biggest challenges in using ML. Training the systems is complicated because extensive knowledge and computing power are needed, but DT does have the resources in their large organization to overcome these challenges.

According to the DT Privacy expert, the main opportunities to implement ML in the DT IoT are identifying private data patterns in the payload IoT. This research area has never been looked at in the current setup of the DT IoT and will be a useful discovery to the company to enhance their privacy-by-design in their IoT and CoT offerings.

4.3 Evaluation of the RB Test Results

The implementation of RB with the same training data shows a straightforward selection of the targeted PII condition: the location data. Since it is a precise calculation or filtering, obtaining a result seems immediate compared to the ML approach. However, the RB conditions did not include all PII related fragment names. The logic in the RB conditions needs to be updated and maintained if the DT CoT implements the RB approach. Moreover, the RB approach should be used with the real IoT payload data, either at the SmartREST API stage or the REST API stage.

The DT AI expert also agreed to the RB approach to detect PII data in the incoming IoT payload, which requires further validation from the real data dump. In his opinion, the main driving factor for implementing the RB method in DT organizations is the awareness of the risk of private data leaks. The higher the risk viewed by the organization, the higher the chances to implement the measure of PII detection. This opinion is also valid for adopting the ML method for detecting PII data.

The DT Privacy expert thought the RB method would be good for potential PII data which have an exact format or structure described in a rule. These rules are easy to understand for the personnel in charge, who evaluate the RB method’s results and classify the breaches. However, he also thought that the main obstacle is that this method does not identify new PII on its own.

4.4 Recommendation to DT

Both the ML and RB approach have their advantages and disadvantages, as discussed in the evaluation sections of this chapter. Regarding a recommendation for the adoption in DT, it is much easier to approach data leak events from an RB point of view for short-to-medium-term solutions. The algorithm for RB is straightforward, provided that we can identify in advance the exact format of PII that we are going to detect.

ML approach might pay off in the long run. Maintaining RB rules are time-consuming and prone to mistakes, especially if the new types of private data emerge as the technology advances. Maintenance of the ML approach will also be demanding; however, we can always increase the complexity of what the ML model should learn. For example, the ML model can be adapted to learn unstructured data, which the RB method cannot handle, as it caters only for the structured data and the exact conditions. Another example that took a similar approach is the Security Information and Event Management (SIEM) system; the traditional SIEM system is ruled based and full of complexity that has to be maintained by network security professionals. Recent research proposed to replace the traditional SIEM system with the ML SIEM system, using a deep learning approach (Azmi Bin Mustafa Sulaiman et al., 2021).

DT has its own AI organization and enterprise-level AI capabilities and tools to test and train the ML model. Although it may take time to collaborate between different organizations within DT, the company will benefit from streamlining the processes. In addition, the ML model can be adapted to the new type of PII data just by re-training and re-learning the model with different classifiers.

The DT Privacy expert considered the research topic relevant to the company. Detecting the private data leaks would give the data controller an early indication to implement appropriate privacy measures.

IoT data is also considered big data when looking at the data characteristics that are high in volume (volume), high in speed (velocity), and highly heterogeneous (variety) (Plaza, 2020). There are additional characteristics added to big data, which are veracity and value. Veracity refers to the quality of data produced by the IoT devices, while value refers to the extracted benefits from the big IoT data. Considering these characteristics of big data, the question remains—which approach is better for detecting the PII the IoT data payload: RB r ML? Without using real data with RB or ML method, this study cannot state confidently whether the labelling of the private data impacts the performance of IoT devices and the analytic engine of Cumulocity.

4.5 Chapter Summary

In this chapter, the results of the ML and RB implementations were discussed and explained. This chapter also discussed the opinion of both DT AI and DT Privacy experts in implementing the ML or RB approach for detecting PII data. The next chapter summarises the ML and RB discovery throughout this study.

Chapter 5. Conclusions

5.1 Chapter Summary

This chapter aims to discuss the conclusions taken from the theoretical observation of ML and RB approaches as per the research questions.

5.2 What Has the Study Achieved So Far?

This study is tailored to a Deutsche Telekom (DT)–specific environment and their IoT and CoT platform designs. This study started with the need to bridge the DT IoT and CoT gap to identify private data in the incoming IoT data stream into the CoT. The incoming IoT payload format was analysed based on DT CoT documentations and the detailed Cumulocity documentation. One example of PII data was selected to be the reference for the study. The next step is to use either the ML or the RB approach to label this identified PII data from the incoming payload. Based on the outcome, this study will recommend the most effective PII detection mechanism, which the company can validate further with their comprehensive enterprise AI and RB tools.

The study indicated that it could detect PII data via the ML approach using the KNIME data analytics platform. Furthermore, the study demonstrated that the RB approach led to a similar outcome. For both methods, we need to continue studying and examining the results with real data from the IoT data payload to further confirm the results' accuracy.

Ideally, the next step would be to apply the solution presented in this study to the enterprise environment and compare the outcomes.

5.3 Differences between This Research and Other Similar Studies

This research utilised a different approach compared with similar studies. The ML algorithm made use of the KNIME analytics platform for implementation. The RB approach explained data flow and the regular expression tool to detect PII.

In contrast, similar research on private data chose a specific IoT use case and a specific objective. Kaneen et al. (2020) offered a methodology to check GDPR compliance with the implementation of remote patient monitoring. However, the present study aimed not to check whether DT CoT complied with GDPR but to detect private data leaks in the incoming data payload.

5.4 Contribution to the Academic World

The research topic of this thesis is of high interest in the specialised scientific domain of cybersecurity. The study utilised an existing algorithm to achieve the desired result of detecting PII data. Knowing which PII data is in the incoming data payload is important. PII detection is the first step towards a larger plan to protect personal data.

The next step is to label PII data to trigger a higher security protection level. Subsequently, we can create private data monitoring events that give the monitoring systems and responsible personnel real-time information.

5.5 Business Application for Deutsche Telekom

Regarding the business applicability of this study, as the DT AI expert mentioned, the first step towards implementing the results of this research is to take the PII detection algorithm to the organization, using their specific tools.

Different departments in the company should collaborate with the AI department to invest time and resources for a common objective: to protect private data. The advantage of this approach is to utilise the current tooling and workforce capacity that the AI department uses. Integration with Security Operations Centre (SOC), Security Information and Event Management (SIEM), and Deutsche Telekom Honeypot should also be part of the holistic approach maximise the benefit for the company.

In another approach, there is already an ML text-processing solution from DT called SemaSuite, offered by a subsidiary of DT, Telekom Multimedia Solution (Telekom MMS). With the collaboration between DT AI and Telekom MMS, implementing the detection of PII events in the IoT data stream could lead to much faster results.

The next step after detecting private data is to raise security attention to it. For example, the labelled data should be pseudonymised to strip off the link to their global identifier and personal data information while still maintaining the applicability for the data analytics tools to perform their analysis.

Apart from pseudonymisation, the company can move to the next stage, building an event system to monitor private data leakage in case of unauthorised disclosure of private data. This measure will strengthen the ‘confidentiality’ part of the Confidentiality-Integrity-Availability (C-I-A) triad (Rizvi et al., 2020) in the products and services offered by DT.

5.6 Recommendations for Further Research

Further work in this area in a similar research program would serve to fine-tune the algorithm for ML. The next research program can choose deep neural learning for ML natural language processing, for instance, Bidirectional Encoder Representations from Transformers (BERT) classifier to enhance the ML outcome. In addition, the next study should, as much as possible, get real IoT data from the IoT payload and compute all possible PII data in the input file of the ML pipeline. Using the real IoT data, one could compute the processing time and compare the results between different models and classifiers.

APPENDIX 1: Interview Questions – Privacy Expert

Interview Questions

Interview Details

Company Name: _____ Date: _____ Time: _____

Expert Name: _____

Expert Title: _____

Working Experience _____

Privacy experience: _____

Questions to Ask the Expert

Question #1: This thesis aims to detect private information from the incoming IoT data payload to the Cumulocity IoT engine in the Cloud of Things (CoT). Please give your opinion on the research topic. Is this research relevant for private data security in the IoT world?

Notes: _____

Question #2: The study evaluates the machine learning (ML) method to detect private data in the incoming IoT payload. Using a phyton script, we generated the input file mimicking the IoT data payload format. For this study, we only select one type of PII data, which is the location data. The files without location data information were grouped in the negative document class, while the positive document class contained location data. Before the data mining phase, these files went through text processing analysis, data transformation, and data manipulation. We trained the data using a tree-based machine learning classifier and then performed predictions on it. The result showed that the ML model could detect personally identifiable information (PII) (in this case, location data) amongst all other messages in the JSON files randomly generated IoT payload data. We repeated the ML training with a different ML classifier and achieved a similar result.

Do you think the ML algorithm approach is fit for the purpose? Is it feasible to deploy it in the Deutsche Telekom environment?

Notes: _____

Question #3a: What do you think are the main opportunities in implementing an ML-based approach to detect PII leaks in Deutsche Telekom?

Notes:

Question #3b: What do you think are the main obstacles in implementing an ML-based approach to detect PII leaks in Deutsche Telekom?

Notes:

Question #4: The study also explores the Rule-Based method to select private data in the incoming IoT payload. This study used the same input files from the ML method, which contains a mixture of PII and non-PII data. The Rule-Based method just filtered out the location data directly using the regular expressions formula.

Do you think that this RB workflow is fit for the purpose? Is this RB method feasible to be deployed in the Deutsche Telekom environment?

Notes:

Question #5a: What do you think are the main obstacles to implementing an RB-based approach to detect personally identifiable information (PII) leaks in the Deutsche Telekom setup?

Notes:

Question #5b: What do you think are the main opportunities to implementing an RB-based approach to detect personally identifiable information (PII) leaks in the Deutsche Telekom setup?

Notes:

Question #6: Given the ML and RB analysis results, which approach would you prefer to implement in DT?

Notes:

Question #7: In your opinion, if we can detect private data at the Cumulocity IoT engine of the CoT platform, can the company set up a private data events alert as the next stage to trigger some relevant actions such as informing the CoT data owner or creating pseudonymisation options for the PII data?

Notes:

Question #8: In your opinion, will private data alerts enhance Deutsche Telekom's 'Privacy-by-Design' incorporated in their IoT products and services? Can this solution assist the company in acting proactively in possible data breaches?

Notes:

Question #9: Do you have any other comments or suggestions about this research and its applicability in the context of DT?

Notes:

Additional Notes

Enter Additional Notes.

APPENDIX 2: Interview Questions – AI Expert

Interview Questions

Interview Details

Company Name: _____ Date: _____ Time: _____

Expert Name: _____

Expert Title: _____

Working Experience _____

AI experience: _____

Questions to Ask the Expert

Question #1: The study evaluates the machine learning (ML) method to detect private data in the incoming IoT data payload. We have designed a conceptual ML pipeline. Is this pipeline fit for the purpose? Is it feasible to deploy it in the Deutsche Telekom environment? Is this pipeline correctly designed? Do you see a way to improve the pipeline?

Notes: _____

Question #2: What do you think are the main opportunities/obstacles to implementing an ML-based approach to detect personally identifiable information (PII) leaks in Deutsche Telekom?

Notes: _____

Question #3: The study also explores the Rule-Based method to detect private data in the incoming IoT data payload. Do you think that this RB workflow is fit for the purpose? Is this RB method feasible to be deployed in the Deutsche Telekom environment?

Notes: _____

Question #4: What do you think are the main obstacles/opportunities to implementing an RB-based approach to detect personally identifiable information (PII) leaks in the Deutsche Telekom setup?

Notes:

Question #5: This thesis aims to detect private information from the incoming IoT data payload to the Cumulocity IoT engine in the Cloud of Things (CoT). Please give your opinion on the research topic. Do you think it is an interesting research?

Notes:

Question #6: In your opinion, if we can detect private data leakage at the Cumulocity IoT engine, can the company set up a private data event alert as the next stage?

Notes:

Question #7: In your opinion, will private data alerts assist the company in acting proactively in data breaches?

Notes:

Question #8: Do you have any other comments or suggestions about this research and its applicability in the context of DT?

Notes:

Additional Notes

Enter Additional Notes.

References

Aazam, M., Khan, I., Alsaffar, A. A., & Huh, E. N. (2014). Cloud of Things: Integrating Internet of Things and cloud computing and the issues involved. *Proceedings of 2014 11th International Bhurban Conference on Applied Sciences & Technology (IBCAST) Islamabad, Pakistan, 14th - 18th January, 2014*, 414–419. https://doi.org/10.1109/ibcast.2014.6778179

Abomhara, M., & Køien, G. M. (2015). Cyber Security and the Internet of Things: Vulnerabilities, Threats, Intruders and Attacks. *Journal of Cyber Security and Mobility*, *4*(1), 65–88. https://doi.org/10.13052/jcsm2245-1439.414

Abraham, A. (2005). Rule-Based Expert Systems. In P. H. Sydenham & Ri. Thorn (Eds.), Handbook of Measuring System Design (pp. 910–918). John Wiley & Sons. https://doi.org/10.1002/9780470400777.ch5

Alhaidari, F., Rahman, A., & Zagrouba, R. (2020). Cloud of Things: architecture, applications and challenges. *Journal of Ambient Intelligence and Humanized Computing*. Published. https://doi.org/10.1007/s12652-020-02448-3

Ambimat, A. (2021, March 25). IOT and Enterprise Cybersecurity platform. Fido enabled MFA for Enterprise. " blog archive MQTT-Sn – lowering the cost of IOT at scale: Ambisecure- IOT and Enterprise Cybersecurity platform. Fido enabled MFA for Enterprise. AmbiSecure. Retrieved December 18, 2021, from https://ambisecure.ambimat.com/mqtt-sn-lowering-the-cost-of-iot-at-scale/

Ashari, R. (2021). *Internet of things (IoT) Cloud of Things (CoT)*. Leiden University.

Azmi Bin Mustafa Sulaiman, M., Adib Khairuddin, M., Rizal Mohd Isa, M., Nazri Ismail, M., Afizi Mohd Shukran, M., & Abu Bakar Sajak, A. (2021). SIEM Network Behaviour Monitoring Framework using Deep Learning Approach for Campus Network Infrastructure. *International Journal of Electrical and Computer Engineering Systems*, *12*, 9–21. https://doi.org/10.32985/ijeces.12.si.2

Badii, C., Bellini, P., Difino, A., & Nesi, P. (2020). Smart City IoT Platform Respecting GDPR Privacy and Security Aspects. *IEEE Access*, *8*, 23601–23623. https://doi.org/10.1109/access.2020.2968741

Barnaghi, P., Sheth, A., & Henson, C. (2013). From Data to Actionable Knowledge: Big Data Challenges in the Web of Things [Guest Editors' Introduction]. *IEEE Intelligent Systems*, *28*(6), 6–11. https://doi.org/10.1109/mis.2013.142

Barr-Kumarakulasinghe, C., Boon-Kwee, N., & Chan-Yuan, W. (2021). Governing the progress of internet-of-things: Ambivalence in the quest of technology exploitation and user rights protection. *Technology in Society*, *64*, 101463. https://doi.org/10.1016/j.techsoc.2020.101463

Cerbo, F. Di, & Trabelsi, S. (2018). Towards Personal Data Identification and Anonymization Using Machine Learning Techniques. New Trends in Databases and Information Systems. 118–126. https://doi.org/10.1007/978-3-030-00063-9

Chen, M., Miao, Y., Hao, Y., & Hwang, K. (2017). Narrow Band Internet of Things. *IEEE Access*, *5*, 20557–20577. https://doi.org/10.1109/access.2017.2751586

Cavoukian, A. (2012). Privacy by Design [Leading Edge]. *IEEE Technology and Society Magazine*, *31*(4), 18–19. https://doi.org/10.1109/mts.2012.2225459

Cumulocity. (2021). *Cumulocity IoT Guides*. Cumulocity IoT. Retrieved October 31, 2021, from https://cumulocity.com/guides/welcome/intro-documentation/

D'Acquisto, G., Domingo-Ferrer, J., Kikiras, P., Torra, V., de Montjoye, Y.-A., & Bourka, A. (2015). Privacy by design in big data: An overview of privacy enhancing technologies in the era of big data analytics (December Issue). https://doi.org/10.2824/641480

Deutsche Telekom. (2021, December 13). *Internet of Things: Deutsche Telekom's IoT Offering | IoT Telekom*. Retrieved 21–10-24, from https://iot.telekom.com/en

Deutsche Telekom AG. (2021). Data privacy information.

Deutsche Telekom Group. (2020). System description for it/nt system Cloud of Things.

Deutsche Telekom Group. (2021). Authorization_Concept_for_CoT_en-3.1_final.

Dimitrov, D. V. (2016). Medical Internet of Things and Big Data in Healthcare. *Healthcare Informatics Research*, *22*(3), 156–163. https://doi.org/10.4258/hir.2016.22.3.156

El Naqa, I., Li, R., & Murphy, M. J. (2015). *Machine Learning in Radiation Oncology: Theory and Applications* (2015th ed.). Springer.

Emami-Naeini, P., Bhagavatula, S., Habib, H., Degeling, M., Bauer, L., Cranor, L. F., & Sadeh, N. (2019). Privacy expectations and preferences in an IoT world. *Proceedings of the 13th Symposium on Usable Privacy and Security, SOUPS 2017*. Soups. 399–412.

Fathy, Y., Barnaghi, P., & Tafazolli, R. (2018). Large-Scale Indexing, Discovery, and Ranking for the Internet of Things (IoT). *ACM Computing Surveys*, *51*(2), 1–53. https://doi.org/10.1145/3154525

Galaz, V., Centeno, M. A., Callahan, P. W., Causevic, A., Patterson, T., Brass, I., Baum, S., Farber, D., Fischer, J., Garcia, D., McPhearson, T., Jimenez, D., King, B., Larcey, P., & Levy, K. (2021). Artificial intelligence, systemic risks, and sustainability. *Technology in Society*, *67*, 101741. https://doi.org/10.1016/j.techsoc.2021.101741

Hintze, M., & LaFever, G. (2017). Meeting Upcoming GDPR Requirements While Maximizing the Full Value of Data Analytics. *SSRN Electronic Journal*. Published. https://doi.org/10.2139/ssrn.2927540

Hoepman, J. H. (2014). Privacy Design Strategies. *ICT Systems Security and Privacy Protection*, 446–459. https://doi.org/10.1007/978-3-642-55415-5_38

Jeong, H. J., Lee, H. J., & Moon, S. M. (2017). Cloud-based machine learning for IoT devices with better privacy. *Proceedings of the Thirteenth ACM International Conference on Embedded Software 2017 Companion - EMSOFT '17*. Published. https://doi.org/10.1145/3125503.3125626

Jiang, L., & Li, C. (2011). Scaling Up the Accuracy of Decision-Tree Classifiers: A Naive-Bayes Combination. *Journal of Computers*, *6*(7). https://doi.org/10.4304/jcp.6.7.1325-1331

Kamin, D. A. (2017). *Exploring Security, Privacy, and Reliability Strategies to Enable the Adoption of IoT*. Walden University.

Kaneen, C. K., & Petrakis, E. G. (2020). Towards evaluating GDPR compliance in IoT applications. *Procedia Computer Science*, *176*, 2989–2998. https://doi.org/10.1016/j.procs.2020.09.204

Kim, S. H., & Kim, D. (2015). Enabling Multi-Tenancy via Middleware-Level Virtualization with Organization Management in the Cloud of Things. *IEEE Transactions on Services Computing*, *8*(6), 971–984. https://doi.org/10.1109/tsc.2014.2355828

Kishor, A., & Chakraborty, C. (2021). Artificial Intelligence and Internet of Things Based Healthcare 4.0 Monitoring System. *Wireless Personal Communications*. Published. https://doi.org/10.1007/s11277-021-08708-5

KNIME. (2021). *Open for Innovation*. Retrieved December 5, 2021, from https://www.knime.com/

Hon, W. K., Millard, C., & Walden, I. (2011). The problem of "personal data" in cloud computing: what information is regulated?--the cloud of unknowing. *International Data Privacy Law*, *1*(4), 211–228. https://doi.org/10.1093/idpl/ipr018

Le, V. H., den Hartog, J., & Zannone, N. (2018). Security and privacy for innovative automotive applications: A survey. *Computer Communications*, *132*, 17–41. https://doi.org/10.1016/j.comcom.2018.09.010

Li, Y., Krishnamurthy, R., Raghavan, S., Vaithyanathan, S., & Jagadish, H. V. (2008). Regular expression learning for information extraction. *Proceedings of the Conference on Empirical Methods in Natural Language Processing - EMNLP '08*. Published. https://doi.org/10.3115/1613715.1613719

Lukács, A. (2017). *What Is Privacy? The history and definition of privacy.* Tavaszi Szél 2016 Tanulmánykötet I, Budapest. 15-17 April. 256–265. http://publicatio.bibl.u-szeged.hu/10794/7/3188699.pdf

Machina Research. (2016). Global Internet of Things Market To Grow To 27 Billion Devices, Generating USD3 Trillion Revenue in 2025. Machina Research, Gartner IoT Forecast Tool, 2014–2015. https://machinaresearch.com/news/press-release-global-internet-of-things-market-to-grow-to-27-billion-devices-generating-usd3-trillion-revenue-in-2025/

Makkar, A., Kumar, N., Sama, L., Mishra, S., & Samdani, Y. (2020). An Intelligent Phishing Detection Scheme Using Machine Learning. *Advances in Intelligent Systems and Computing*, 151–165. https://doi.org/10.1007/978-981-15-8061-1_13

Malandrino, D., Petta, A., Scarano, V., Serra, L., Spinelli, R., & Krishnamurthy, B. (2013). Privacy awareness about information leakage. *Proceedings of the 12th ACM Workshop on Workshop on Privacy in the Electronic Society*. Published. https://doi.org/10.1145/2517840.2517868

Manyika, J., Chui, M., Bisson, P., Woetzel, J., Dobbs, R., Bughin, J., & Aharon, D. (2015). The Internet of Things: Mapping the Value Beyond the hype. McKinsey Global Institute, 1–162. https://www.mckinsey.com/~/media/McKinsey/Industries/Technology Media and Telecommunications/High Tech/Our Insights/The Internet of Things The value of digitizing the physical world/The-Internet-of-things-Mapping-the-value-beyond-the-hype.pdf

McCallister, E., Grance, T., & Kent, K. (2010). Guide to protecting the confidentiality of personally identifiable information (PII). Special Publication 800-122 Guide, 1–59. https://csrc.nist.gov/publications/detail/sp/800-122/final

Mehmood, Y., Görg, C., Muehleisen, M., & Timm-Giel, A. (2015). Mobile M2M communication architectures, upcoming challenges, applications, and future directions. *Eurasip Journal on Wireless Communications and Networking*, 2015(1), 1–37.

Microsoft. (2021, August 12). *Azure IoT Hub device-to-cloud options*. Microsoft Docs. Retrieved December 12, 2021, from https://docs.microsoft.com/en-us/azure/iot-hub/iot-hub-devguide-d2c-guidance

Perera, C., McCormick, C., Bandara, A. K., Price, B. A., & Nuseibeh, B. (2016). Privacy-by-Design Framework for Assessing Internet of Things Applications and Platforms. *Proceedings of the 6th International Conference on the Internet of Things*. Published. https://doi.org/10.1145/2991561.2991566

Pigni, F., Bartosiak, M., Piccoli, G., & Ives, B. (2018). Targeting Target with a 100 million dollar data breach. *Journal of Information Technology Teaching Cases*, *8*(1), 9–23. https://doi.org/10.1057/s41266-017-0028-0

Plaza, D. J. C. (2020). *Heterogeneous Data Processing In The Internet Of Things*. Universidad de Cádiz. https://www.researchgate.net/profile/David-Corral-

Plaza/publication/350401020_Processing_Heterogeneous_Data_in_the_Internet_of_Things/links/605dc454458515e8347083af/Processing-Heterogeneous-Data-in-the-Internet-of-Things.pdf

Qu, Z., Zhang, G., Cao, H., & Xie, J. (2017). LEO Satellite Constellation for Internet of Things. *IEEE Access*, *5*, 18391–18401. https://doi.org/10.1109/access.2017.2735988

Renner, P. (2021). *Securing Internet of Things (IoT) Devices that Interact with Personal Information*. (Vol. 4, Issue 1). Utica College.

Rizvi, S., Pipetti, R., McIntyre, N., Todd, J., & Williams, I. (2020). Threat model for securing internet of things (IoT) network at device-level. *Internet of Things*, *11*, 100240. https://doi.org/10.1016/j.iot.2020.100240

Sahmim, S., & Gharsellaoui, H. (2017). Privacy and Security in Internet-based Computing: Cloud Computing, Internet of Things, Cloud of Things: a review. *Procedia Computer Science*, *112*, 1516–1522. https://doi.org/10.1016/j.procs.2017.08.050

Sink, M. (2001). The Use of Honeypots and Packet Sniffers for Intrusion Detection. *GIAC Security Essentials*. Version 1.2b. SANS Institute. April 15, 2001. 1–7.

Solove, D. J. (2008). *Understanding Privacy*. Harvard University Press. May 2008. GWU Legal Studies Research Paper No. 420. Available at SSRN: https://ssrn.com/abstract=1127888

Topaz, M., Murga, L., Gaddis, K. M., McDonald, M. V., Bar-Bachar, O., Goldberg, Y., & Bowles, K. H. (2019). Mining fall-related information in clinical notes: Comparison of rule-based and novel word embedding-based machine learning approaches. *Journal of Biomedical Informatics*, *90*, 103103. https://doi.org/10.1016/j.jbi.2019.103103

van den Berg, J., van Zoggel, J., Snels, M., van Leeuwen, M., Boekee, S., Koppen, L., van den Berg, B., de Bos, A., & van der Lubbe, JCA. (2015). On (the emergence of) cyber security science and its challenges for cyber security education. In E. Luijff (Ed.), *Proceedings of the NATO IST-122 Cyber Security Science and Engineering Symposium, Tallinn, Estonia, October 13–14 2014*. NATO Science and Technology Organization.

Verwer, S. (2021). Introduction to Machine Learning and Data Mining. Lecture Slides, TU Delft.

Vojković, G., Milenković, M., & Katulić, T. (2020). IoT and Smart Home Data Breach Risks from the Perspective of Data Protection and Information Security Law. *Business Systems Research Journal*, *11*(3), 167–185. https://doi.org/10.2478/bsrj-2020-0033

Waheed, N., He, X., Ikram, M., Usman, M., Hashmi, S. S., & Usman, M. (2021). Security and Privacy in IoT Using Machine Learning and Blockchain. *ACM Computing Surveys*, *53*(6), 1–37. https://doi.org/10.1145/3417987

Warren, S. D., & Brandeis, L. D. (1890). The Right to Privacy. *Harvard Law Review*, *4*(5), 193. https://doi.org/10.2307/1321160

Wolford, B. (2019, February 13). *What are the GDPR Fines?* GDPR.Eu. Retrieved October 24, 2021, from https://gdpr.eu/fines/

Yun, J., Ahn, I. Y., Choi, S. C., & Kim, J. (2016). TTEO (Things Talk to Each Other): Programming Smart Spaces Based on IoT Systems. *Sensors*, *16*(4), 467. https://doi.org/10.3390/s16040467

Zhou, L., Pan, S., Wang, J., & Vasilakos, A. V. (2017). Machine learning on big data: Opportunities and challenges. *Neurocomputing*, *237*, 350–361. https://doi.org/10.1016/j.neucom.2017.01.026