TECNOLOGIAS
E ARQUITETURA

Department of Information Science and Technology

# Waste Management in Smart Cities: Optimization of Waste Container's Capacity using Fixed-Frequency Collection

José Tiago Barata Pereira da Costa

A Dissertation presented in partial fulfillment of the Requirements

for the Degree of

**Master in Telecommunications and Computer Engineering**

**Supervisor:**

Prof. Dr. João Carlos Ferreira, Assistant Professor

ISCTE-IUL

September, 2020

## Acknowledgments

In this section I would like to take the opportunity to express my gratitude towards the people that helped me through the process of developing and writing this dissertation.

A special thanks to my family, my Parents Gabriela and Nuno, to my Grandparents Eduardo and Antonieta, and to my Girlfriend Ana, that gave me motivation and unconditional support through my studies and always encouraged me for following my goals.

I would like to thank my supervisor, Prof. Dr. João Carlos Ferreira, for his support, guidance and patience demonstrated over the last year and half.

I am thankful to ISCTE – Instituto Universitário de Lisboa and Department of Information Science and Technology for the opportunity and the conditions which allowed me to complete my education.

## Resumo

Um dos principais problemas das sociedades contemporâneas é o controlo do fluxo de produção e remoção dos resíduos sólidos urbanos, devido à massificação contínua das zonas urbanas.

Esta dissertação incide sobre a necessidade de reduzir o impacto da atividade humana no meio ambiente, através da gestão dos resíduos sólidos urbanos. Neste trabalho é descrito como a evolução tecnológica pode conduzir a um desenvolvimento urbano mais sustentável, através de um planeamento mais eficiente, redução dos custos logísticos e das emissões poluentes.

O trabalho utiliza dados de um produto 360Waste da Evox Technologies, uma empresa de um antigo aluno do ISCTE, que opera em Castelo Branco. Esta empresa é especializada na criação de uma solução integrada para a recolha eficiente de resíduos urbanos. Esta solução é composta por sensores de leitura volumétrica, com base na tecnologia LoRaWAN. Estes sensores são instalados em contentores de resíduos sólidos urbanos, que estão constantemente a enviar dados para uma *Gateway* LoRaWAN, cada vez que um indivíduo abre o contentor.

Com base nos dados recolhidos dos sensores, o desafio deste trabalho de investigação será desenvolver uma solução para optimizar a gestão dos resíduos sólidos urbanos, definindo um sistema de recolha uniforme e utilizando tecnologias conhecidas como Data Sciences e Machine Learning.

**Palavras-Chave:** Capacidade-Frequência, Logística, Transportação, Recolha de Resíduos, Big Data, *Data Mining*, Machine Learning, Internet das Coisas, LoRa, LoRaWAN, IoT, Smart Cities.

## Abstract

One of the main problems of modern societies is the control of the production flow and removal of urban solid waste, due to the continuous massification of urban areas.

This dissertation focuses on the need to reduce the impact of human activity over the environment through the management of urban solid waste. It describes how technological advancements can lead to an increase in the sustainability of urban development, through more efficient planning and reduction of logistics costs and pollution emissions.

The work uses data of the product 360Waste from Evox Technologies, a company of a former student of ISCTE, operating in Castelo Branco. This company is specialized in creating an integrated solution for efficient collection of urban waste. This solution is composed of volumetric reading sensors, based on LoRaWAN technology. These sensors are installed in urban solid waste containers, which are always sending data to a LoRaWAN gateway, every time an individual opens the container.

Based on the data collected from the sensors, the research work challenge will be to develop a solution to optimize the management of urban solid waste, by defining a uniform collection system and using technologies known as Data Sciences and Machine Learning.

**Keywords:** Frequency-Capacity, Logistics, Transportation, Waste Collection, Big Data, *Data Mining*, Machine Learning, Internet of Things, LoRa, LoRaWAN, IoT, Smart Cities.

Content

**Acknowledgments** ..... **i**

**Resumo** ..... **ii**

**Abstract** ..... **iv**

**List of Figures** ..... **viii**

**List of Tables** ..... **x**

**List of Acronyms** ..... **xi**

**Chapter 1 - Introduction** ..... **1**

1.1. Context and Motivation ..... 1

1.2. Objectives ..... 2

1.3. Thesis Outline ..... 3

**Chapter 2 - State of Art** ..... **4**

2.1. Recycling and *Reverse Logistics* ..... 4

2.2. Vehicle Routing Problems (VRP) ..... 5

2.3. Waste Collection Management Constrains ..... 7

2.4. *Smart* Sensors ..... 9

2.4.1 LoRa and LoRaWAN ..... 9

2.4.2 LoRaWan ..... 10

2.5. Waste measure sensor ..... 12

2.6. Data Analyses ..... 13

**Chapter 3 - Methodology** ..... **16**

3.1. Business Understanding ..... 17

3.2. Data Understanding ..... 18

3.3. Data Preparation ..... 19

3.4. Data exploration ..... 20

3.5. Deposits and Collections ..... 24

3.6. Data Correlation ..... 27

3.7. Major Findings ..... 28

**Chapter 4 - Model Optimization ..... 30**

4.1 Model Formulations for Optimizing Collection Frequency ..... 30

4.2 Comparison of the proposed models ..... 30

4.3 Application of the selected model ..... 33

**Chapter 5 – Prediction Process ..... 37**

5.1. Determine the objective ..... 37

5.2. Choice of a learning problem and its type ..... 38

5.3. Data preparation ..... 39

5.4. Evaluate algorithms ..... 40

5.4.1 Decision Tree score ..... 41

5.4.2 Random Forest Score ..... 42

5.4.3 K-Nearest Neighbors score ..... 44

5.4.4 Support Vector Machine score ..... 45

**Chapter 6 - Conclusions and Future Work ..... 47**

6.1 Conclusions ..... 47

6.2 Future work ..... 49

**References ..... 51**

## List of Figures

Figure 1- State of Art summary ..... 4

Figure 2 - Reverse Logistics Process[11]. ..... 5

Figure 3 - The LoRaWAN (MAC) protocol stack implemented on top of LoRa modulation (PHY) [27]. ..... 10

Figure 4- LoRaWan Network with a star-of-stars topology. ..... 11

Figure 5- 360Waste sensor [28]. ..... 12

Figure 6- Current solution of the waste management system, with route optimization based on capacity available [28]. ..... 17

Figure 7- Standard waste containers (ids: 48843, 52910) and surface waste containers (ids: 44263, 44966, 50419). ..... 18

Figure 8. Container's streets locations with a perspective view. ..... 19

Figure 9. Container's life cycle. ..... 21

Figure 10. The average volume of waste in the containers by street location, and their according months. ..... 22

Figure 11. Average container's volume in percentage for each day of week. ..... 23

Figure 12. Container's volume along the day. ..... 23

Figure 13. The average volume of litres deposited and collected per day. ..... 25

Figure 14. The average volume of waste collection and deposit, for each month and it is waste collection frequency ..... 26

Figure 15. Pie chart representation of the deposits by the levels of precipitation, temperature, season, and type of day. ..... 27

Figure 16. Model for three times a week. ..... 31

Figure 17. Model for two times a week in summer season. ..... 32

Figure 18. Model for two times a week. ..... 33

Figure 19. Average volume for each street. ..... 34

Figure 20. Average container's volume for each street location. ..... 34

Figure 21. Street Capacity optimization. ..... 35

Figure 22. Average volume for each street with optimization. ..... 36

Figure 23 - Waste volume above 60%..... 38

Figure 24. Machine learning prediction process. ..... 40

Figure 25. Decision Tree Graph ..... 41

Figure 26. Random Forest's Confusion Matrix ..... 44

Figure 27: K-Nearest Neighbors Accuracy Score. ..... 45

Figure 28 - K-Nearest Neighbors F1- Score..... 45

# List of Tables

Table 1 - Decision Tree Accuracy Score. ..... 42

Table 2 - Random Forest score ..... 43

Table 3 – Support Vector Machine Accuracy Score. ..... 46

## List of Acronyms

| | |
|---|---|
| CRISP-DM | Cross Industry Standard Process for *Data Mining* |
| HFFVRP-IF | Heterogeneous Fixed Fleet Vehicle Routing Problem with Intermediate Facilities |
| IoT | Internet of Things |
| ISM | Industrial, Scientific, and Medical |
| KDD | Knowledge Discovery in Databases |
| KNN | K-Nearest Neighbors |
| LoRa | Long Range |
| LoRaWAN | Long Range Wide Area Network |
| LPWAN | Low Power Wide Area Network |
| MCVRP | Multi-Compartment Vehicle Routing Problem |
| ML | Machine Learning |
| PRP | Pollution-Routing Problem |
| SVM | Support Vector Machine |
| TSP | Travelling Salesman Problem |
| VRP | Vehicle Routing Problem |
| VRSP | Vehicle Routing and Scheduling Problem |

# Chapter 1 - Introduction

## 1.1. Context and Motivation

According to United Nations statistics [1], around 4.2 billion people live in cities nowadays. That is, over 55% of the global population lives in urban areas [2], and the trend is for this number to continue to grow by 1.8% every year. This means that the future is urban for most of the people in the world, and the solutions to some of Humanity's most outstanding issues, such as poverty, climate change, healthcare, and education, must be found in city life.

Indeed, cities occupy just 3% of the Earth's land but already account for 60% to 80% of energy consumption, according to the United Nations data [1]. Waste production is also a fast-growing problem of modern societies, particularly in growing urban regions. The accumulation of solid waste in urban areas is becoming a great concern. Around 1.7–1.9 billion metric tons of municipal solid waste is produced every year worldwide [3], and the forecast is for an increase of 70% by 2025 [4].

According to the official website of European Commission [5], in Europe, each person is expected to yearly produce six tons of waste materials used in the daily life, which would result in environmental pollution and risks to human health if not properly managed. For instance, several projections and strategies should be approached for efficient waste management, such as building a structured process for the waste disposal and maximizing the recycling of the waste towards making the system as economical and sustainable as possible.

A sustainable development [3] implies concern and articulation between the economic, social and environmental areas in a global economy context. An effective establishment of sustainable and economic development requires a new vision and new forms of action at local, regional, and global levels.

Internet of Things (IoT) devices [6] are an example of how the advance in technology can contribute to sustainable development, improve the quality of citizens' life and increase economic growth. They are a central component of the management infrastructure of what is known as *Smart Cities* [7]. These IoT devices are starting to be used in fields such as waste management, smart traffic management, industrial management, structural health monitoring, security, emergency services, supply chain,

retail, healthcare, and other community services, providing real-time information of the environmental conditions where they are installed.

The amount of data that they produce leads to a stringent requirement for *data analysis* and data storage platforms, which are needed to maximize their potential.

This case study comprises of a physical set of waste containers with IoT sensors, which allow each one to report its filling level. The advanced functionality of such a system enables to predict the expected emptying time of a recycling container, i.e., the time when the container's filling level will reach a certain critical value.

With this data collected from the waste containers, some analysis will be carried out in order to understand some patterns and behaviors of the citizens, thus avoiding useless collections, overfilling violation and unloading requirements.

This dissertation describes a possible approach to use IoT sensor data to improve services and save resources. It uses data collected from sensors installed in waste containers of the city of Castelo Branco, Portugal, to perform extensive data analysis methods and algorithms based on Machine Learning (ML) techniques that can be used for such a system efficiently.

## 1.2. Objectives

The purpose of this dissertation is to optimize the capacity of urban solid waste containers in the area of Castelo Branco, Portugal, using data provided by sensors located inside each waste container. These sensors collect data of the volume of residuals inside, each time the container is opened.

To perform the optimization, a uniform system has to be defined. For that, multiple scenarios will be discussed such as the capacity of each waste container, the number of containers per street, the number of collections per week, atmospheric conditions, celebrative days, ( as Christmas day) and the seasons.

With the use of existing software, data analysis techniques were applied to the data to obtain a predictive model of the system – that is, use past information to predict the future behavior of the waste containers and from that derive the schedules to optimize the waste collection.

For example, to predict the next day's containers capacity, it is necessary to extract records of the previous days. Hence, the problem formulation can be done in two stages:

a) Prepare the dataset containing all the variables, and include more in the future, if needed. Analyze the dataset and try to extract *knowledge* with graphics and create correlations.
b) *Machine Learning* model design: Usage of Machine Learning techniques to predict if a certain street location needs to be collected, regarding a uniform waste collection scenario.

## 1.3. Thesis Outline

After this introduction, Chapter 2 presents the Literature review – the State of the Art that summarizes all the related topics. It describes the research done to frame the thesis theme, such as the multiple types of vehicle routing problems in waste management systems; studies the behavior of the solid waste management in urban areas and their constrains; the waste management in *Smart Cities* and a theory review of the LoRa and LoRaWan technology.

Chapter 3 aims to describe the system methodology – Procedure of data cleaning and mining. From that data, exploration techniques were performed using graphical data analyzis in order to find correlations and patterns.

Chapter 4 refers to the proposal and evaluation of different models with a defined uniform collection system. From that, a comparison between these models is made, in order to select the one that can better simulate the filling levels of the waste containers.

Chapter 5 concerns the application of supervised ML algorithms to predict if a certain street (as a set of waste containers) needs to have the waste collection.

Chapter 6 presents the conclusions and future work - Synthesis of the global results of the analysis, summarizes the main contributions and identifies the limitations of this study.

# Chapter 2 - State of Art

This chapter describes the literature review carried out for the proposed work. It introduces the concepts that are essential to this work, namely Municipal Solid Waste logistics management, Vehicle Routing Problems, Smart Waste management using IoT sensors, the sensors technology, and the technology used to perform the data analytics (see Figure 1).

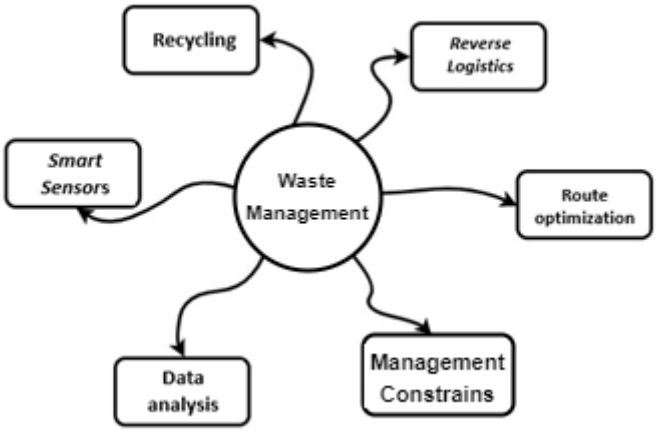


*Figure 1- State of the Art main topics*

## 2.1. Recycling and *Reverse Logistics*

Proper waste management is necessary to handle the increased waste flow in urban areas. Each area has its own specific characteristics and constraints, so a proper waste collection system and treatment has to be specified.

Regarding the treatment of the collected waste, recycling is a common treatment option, as well as the most sustainable and environmentally friendly. The recycling increase leads to lower environmental impact, lower consumption of energy sources and lower economic costs [8].

The statistics from EUROSTAT (2018) [9] shows that in the European Union (including UK), the recycling rate of municipal waste is about 47%. The countries with the highest recycling rates are Germany, with 67.3%, Slovenia, with 58.9%, Austria, with 57.7% and the Netherlands with 55.9%.

However, there are still many countries in the EU that do not recycle or recycle very little, such as Serbia 0.3%, Montenegro 5.5%, Malta 6.5%, Romania 11.5% and so on.

Moreover, in the EU, more than 40% of waste is currently sent to landfills. This is an indicator that there is still a margin to improve the recycling of the municipal solid waste. Another treatment of the collected waste is *Reverse Logistics*. This process consists of reusing products that are no longer required by end-users (see Figure 2). Those products are then transformed into new products with a value in the market [10]. This relies on planning strategical and tactical decisions, such as logistics network design and collection design.


*Figure 2 - Reverse Logistics Process*[11].

## 2.2. Vehicle Routing Problems (VRP)

The Vehicle Routing Problem (VRP) [12] is an extension of the familiar Travelling Salesman Problem. The Travelling Salesman Problem is an NP-hard combinatorial problem, resulting in successive iteration procedures, in order to discover an optimal solution, and differs for each case. Its general formulation is:

*"Given a list of cities and the distances between each pair of cities, what is the shortest possible route that visits each city and returns to the origin city?"* [13].

In VRP, the problem consists in finding an optimal set of routes for a group of vehicles to cross in order to serve a given set of locations, considering multiple constraints. The first mention of a VRP appeared in a paper The Truck Dispatching Problem by George Dantzig and John Ramser in 1959 [14], in which the first algorithmic approach was written and applied to petrol deliveries.

This idea of minimizing the total route cost can also be applied in waste management problems, either for a dynamic or fixed route. Some popular **extensions** of vehicle route problems related to waste management are:

- Pollution Route Problem (PRP).
- Capacitated Vehicle Routing Problem (CVRP).
- Multi-Compartment vehicle routing problem (MCVRP).
- Vehicle Routing and Scheduling Problem (VRSP).
- Heterogeneous Fixed Fleet Vehicle Routing Problem with Intermediate Facilities (HFFVRP-IF).

The PRP measures the travel distances, the number of greenhouse emissions, fuel consumption, number of travel routines and their costs. Paper [15] proposes a mathematical model for the PRP with time window constraints and takes into consideration multiple parameters such as vehicle load, speed and total cost.

The CVRP considers constraints regarding vehicle capacity. It was initially introduced by Dantzig and Ramser (1959) [16], and has received a huge amount of attention in the field. In this context, the CVRP is defined as having a set of waste containers to be served by a fleet of collection trucks, where they start and finish at the waste depot.

Paper [17] proposes a modified particle swarm optimization (PSO) algorithm to determine the best waste collection and route optimization solutions, taking into account the capacity of the trucks, distance, efficiency and cost. In this article, both the efficiency of waste collection and route optimization process is measured using the data obtained from smart waste containers, which are assembled with LoRa custom hardware and its protocol. These sensors were used to measure the waste levels of the containers, and from there, determine a viable route according to the following constraints:

- All vehicles must start and return to a specified depot.
- A waste container may only be visited by one vehicle each time.
- The total capacity of a vehicle must not exceed its maximum.
- A container needs to be emptied as soon as possible after it reaches its predetermined threshold waste level.

The MCVRP extension is similar to the CVRP model, however, in this case the vehicles are equipped with multi-compartment storages, which enables them to store various types of waste without mingling them (co-collection).

Paper [18] presents an MCVRP in the context of glass waste collection. A model with a heuristic approach was proposed for a vehicle with multi-compartment storage, which is able to pick up the different glass types of the suppliers and move them to the nearest depot.

The VRSP is an extension of a VRP with a time horizon constraint. It is a model developed for systems where it is necessary to perform a real-time recalculation of routes and schedules aspects.

Article [19] contains a case study of VRSP in the field of urban environmental management that describes the collection of residuals by a public company that supervises the collection of solid household and street waste in five districts of Hanoi (Vietnam).

To optimize the operations, it was proposed a heuristic procedure applying a version of *Solomon's* model for the construction of the routes and the usage of Or-opt and 2-opt algorithms to improve the routes. It reaches a solution with an improvement in both total cost and number of vehicles utilized. Also, in the article [20], a model for scheduling and routing of the streets of a Swedish city was proposed, where 3300 waste containers were fitted with sensors and wireless equipment for real-time information of each container.

The HFFVRP-IF is a natural generalization of VRP with several vehicle types, each of them classified by its waste collection type, its capacity, fixed cost per distance unit, and a flexible assignment of destination depots. It is designed to cover more practical situations in real-life transportation.

In Paper [21], this model is proposed with multiple constraints, such as starting time periods and site dependencies, i.e. multiple disposal sites.

The results showed that the proposed heuristic model achieved optimal results in small instances and could lead to important savings when properly understood and applied.

## 2.3. Waste Collection Management Constrains

Waste management requires facing many challenging issues according to technical, financial, economic, and social constraints. The state of the art of this dissertation covers

technical and logistical constraints regarding the collection of the residuals. It is necessary to determine the number of times per week that each specific container should be collected, which are often located in the same street with other container types. Some of the common waste container types are domestic waste, paper, plastic, and glass. Other waste types are packaging waste, end-of-life vehicles, batteries, electrical and electronic waste.

Certain containers can also have special necessities, which are denominated in the article [22] as high priority containers that require immediate collection, i.e., schools or hospitals, where an algorithm has been developed to respond to the demand of these containers, which are detected by sensor observations. Several simulations were conducted in order to maximize the performance in terms of economical and quality of the service.

Alongside the collection frequency, it is also necessary to specify the most optimal way to define and schedule a route. Route scheduling can be dynamic or fixed. In the case of dynamic, real-time recalculation of routes and schedules are performed according to the demands and the conditions of the system, such as the road condition. In this literature review (section 2.2) various models and techniques were presented that can be applied to the system optimization. In case of designing a fixed route schedule, a route is delineated according to the shortest path possible from the start point to the endpoint and is defined a fixed collection time routine, using data analytics based on pre-records of data, such as waste generation, city development index, population density and other external factors.

The traffic regulations and the road condition are also a determinant factor in waste management, as the route collection has to be adapted to the conditions of the environment and the terrain, especially in old cities. In [23] the authors proposed a CARP model considering streets that can only be traversed in one direction, streets with narrow-angle of vision, restrict streets crossings and traffic signals. Also, in [21], some road conditions were measured, as the big collector truck couldn’t reach the mountainous terrain and narrow streets.

Waste management also needs to take into account the location of the depots, which is where the residuals collected by the municipals are processed. The number of depots depends on each city’s needs and might receive a variety of elements, such as recycling items, household items, yard waste, electronics and other items for reverse logistics or

disposal. Paper [24] describes a heuristic approach for the definition of the service areas of multiple depots in a reverse logistics network. The system optimization considers objectives related to economic and organizational issues, namely, minimization of the travel costs by the collection vehicles and the pursuit of equity, leading to balance the workload differences among depots.

The number and the type of vehicles available for collection is also an important factor concerning the optimization of routes and collection timetable. The type of vehicle consists in the volume of litres they can store, the type of waste that they are available to collect and other factors, such as the type of energy consumption, which can be electric, oil or gas. There, as a vehicle capable of storing different types of waste, these multi-compartment vehicles are described in article [18], which presents a solution in which MCVRP works in the simultaneous collecting of two or more types of waste without mixing them (co-collection).

## 2.4. *Smart* Sensors

Internet and its applications have become an integral part of today's human lifestyle. After a tremendous increase in demand and necessity, researchers went beyond traditional communication techniques, connecting digital equipment to the internet and to each other, creating the Internet of Things (IoT).

Home automation business and transportation industries are experiencing rapid growth with IoT. The desire to interact with other users and machines is no longer a dream, but a reality.

### 2.4.1 LoRa and LoRaWAN

LoRa is the acronym for Long Range and is a wireless technology where a low powered sender transmits small data packages (0.3kbps to 5.5 kbps) to a receiver over a long distance (up to 10 kilometres). This type of communication offers a novel communication paradigm, which is being used in devices for Smart Cities, as IoT applications. This technology provides a network for low cost, long-range and energy-efficient, which are the requirements to set up the waste containers with LoRa sensor devices.

LoRa uses a radio modulation technique Chirp Spread Spectrum [25], containing a wider band, thus uses the entire channel bandwidth to broadcast a signal which makes it resistant to channel noise, long term relative frequency, Doppler effects and fading. The characteristics of LoRa are based on three basic parameters: Code Rate, Spreading Factor and Bandwidth.

The LoRa devices have very high sensitivities, which leads to high link budgets, so if these devices are used with the line of sight, if the path loss is lower than the radio link budget (LoRa device), a network connection is possible.

According to the LoRa Alliance specifications [26], it operates at the industrial, scientific and medical (ISM) frequency band, which means 169 MHz, 433 MHz and 868 MHz for Europe and 915Mhz for the United States. Since it operates in the ISM band, the devices should respect some rules relative to the duty cycle and Effective Radiated Power.

### 2.4.2 LoRaWan

While LoRa represents the physical layer, enabling the Long-Range communication link, LoRaWAN defines the communication protocol and system architecture network which is an open-source communication protocol defined by the LoRa Alliance [26] consortium (see Figure 3).

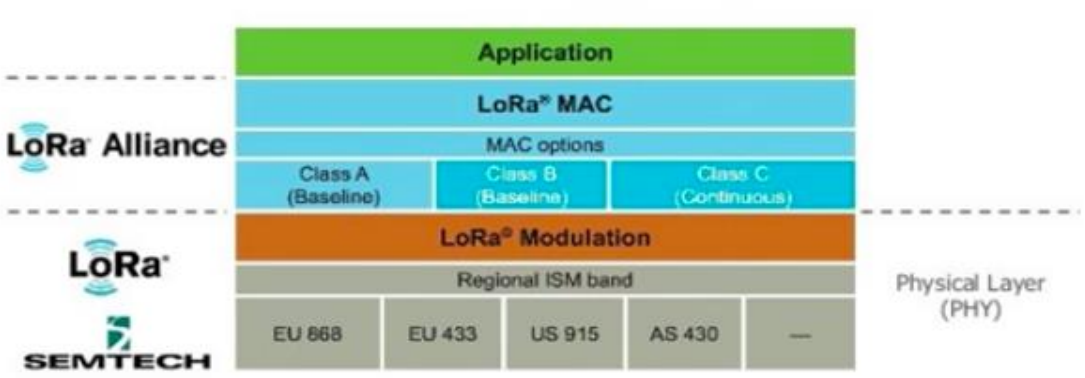


*Figure 3 - The LoRaWAN (MAC) protocol stack implemented on top of LoRa modulation (PHY)* [27].

LoRaWAN communication protocol ensures reliable and secure communication and adds additional headers to the data packets.

LoRaWAN end-devices can be from three different classes:

- **Class A**: Bi-directional end-devices.
- **Class B**: Bi-directional end-devices with scheduled receive slots.
- **Class C**: Bi-directional end-devices with maximal receive slots.

LoRaWan network architecture is deployed in a star-of-stars topology (see Figure 4).

- **End Nodes:** transmit data directly to all gateways within range, using LoRa.
- **Gateways:** relay messages between end-devices and a central network server using IP.

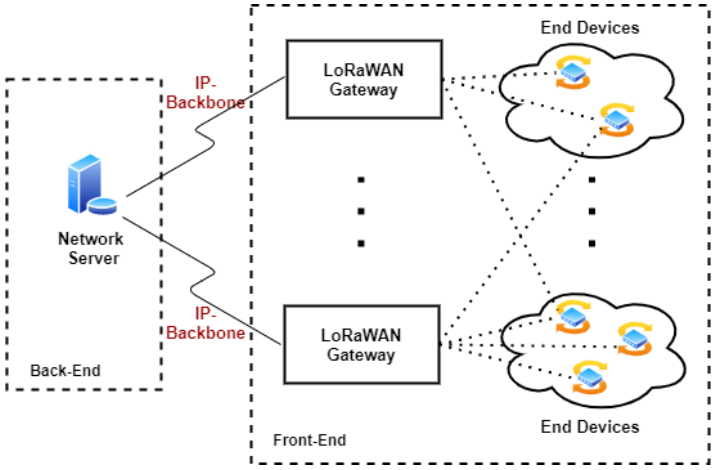


*Figure 4- LoRaWan Network with a star-of-stars topology.*

Also, Figure 4 shows a separation of the architecture in two: front-end and back-end.

- **Front-end:** The communication established between the end-devices and the gateways through LoRa modulation protocol.
- **Back-end:** The network server responsible for the storage receives information from the connection between the gateways and the central server, handled over a backbone IP-based network.

To conclude the Lora and LoRaWAN analysis, the advantages of the usage of this kind of technology in LPWANs are presented.

- ISM frequency band
- Scalable
- Bi-directional communication
- High level of security due to encryption

- Low power
- Low-cost
- Environmentally friendly

### 2.5.Waste measure sensor

Extracting valuable information from data is one of the most important tasks in business organizations. Hence, data from the sensors inside the waste containers is transmitted via LoRa communication to a central database.

Each sensor (see Figure 5) uses ultrasonic technology to measure the volume of the container in real-time, and is compatible with various materials, solids, and liquids. It is energetically independent with an autonomous communication system. As the IoT devices, these sensors are easy to install, robust, resistant and have warning fire system incorporated.

These sensors communicate via LoRa (Long Range) network, which are Low Power Wide Area Network technologies (LPWAN) and is powered by batteries with the autonomy of approximately ten years, ranging from about 30 to 50km in rural areas and 3 to 10 km in urban ones.


*Figure 5- 360Waste sensor* [28].

360Waste sensor datasheet [28]:

- Proprietary polyurethane
- Diameter 120mm.

- Height 45mm.
- Weight 350g.
- Range 18cm to 350cm.
- Ultrasound direction 80º.
- Accuracy of +- 2cm.
- 10 years battery lifetime*.
- Improved lithium battery.
- Communications via GPRS, LoRa, NB-IoT.
- Location by GPS (optional).
- Working temperatures -20º - +80ºC.
- 4 x M5 screw.

*Battery lifetime may vary depending on the sensor update frequencies, network quality and surrounding temperature.

### 2.6. Data Analyses

In principle, data analytics aims to examine raw data to uncover hidden patterns, correlations, and other insights. Waste management services and companies may apply analytics to describe, predict and improve their operation from available monitoring data, statistics, and other pieces of information.

The data is often collected from a datastore server, and when it is significant, it is common to name it as *Big Data*. As a result, this data is often made up of volumes of text, dates, numbers, or complex classes with a classification type.

In this context, the most important data are the datetime and the volume of waste inside the container. Every time that an individual opens the container, the ultra-sound sensor records its filling-level and the datetime of the current moment, which sends it via LoRa to a LoRaWAN gateway, to then redirect it to the server to store in the database.

The purchasing data of a container can be aggregated over several time levels, such as an hour of the day, day of the week, month and season, simultaneously, and can extract both independent time level patterns and inter-relationship patterns among the time levels used [29].

The data can be exported in different formats, such as database table, *csv*, text, excel and others. For the analysis purpose, different types of variables are considered; in terms of machine learning, variables are named *features*. In the knowledge discovery process, the data should be in the proper format. Converting the raw data into the proper format is the initial step of Knowledge Discovery in Database (KDD) process, called pre-processing. In this study, raw data is imported from a *csv* format, and converted, using the Python programing language , into a data frame table, from the *Pandas* library (IDE Jupyter Notebook).

A simple technique to predict the **accuracy** of the model is the application of the If-Then-Else rules. Some Machine Learning algorithms use this technique, as the **Decision Tree**. For example, if a container gets fulfilled every twelve hours very often, probably in the next twelve hours, it will be full again. Or If it is a rainy day, containers do not get full very often.

Paper [30] presents the use of an automated machine learning for detecting the accuracy of the IoT sensors, implemented in the waste containers, were multiple classifications algorithms have been tested: the artificial neural, k-nearest neighbours, logistic regression, support vector machine, decision tree and random forest.

Paper [31] also uses the concepts of IoT sensors in the waste containers, called “Smart Bins”, to collect data in real-time and offering monitoring capability. The good levels of accuracy and efficient data analytics, lead them to an experiment based on the re-schedule of the routes, taking into account the time savings, the fuel consumption and capacity reached of the waste containers. In this experiment, the results recorded from a ten-day trial allowed a significant optimization in fuel and time costs, compared with the previous pre-defined fixed route.

In paper [32] it is addressed the assessment and benchmarking of selective collection schemes in order to improve future waste collection operations. These schemes are based on statistical analysis and monitoring, using three performance indicators: Effective Collection Distance, Effective Collection Time, and Effective Fuel Consumption. Those indicators were analyzed taking into account the type of collection: Drop-off and Street Side collection; the waste type material: Light Packaging, Paper/Cardboard and Glass, and also using descriptive and inferential statistics. In further analyses, they reached into

a generic formula which combines those indicators and their weights to generate different performances according to the stakeholders' interests.

In article [33], in the Helsinki metropolitan area (Finland), data analyses were made from a different perspective. Here information on waste production by the citizens is used for the basis of waste management and collection. The data analytics process was explored using the Knowledge Discovery in Database's (KDD) technique, where the stages of data processing are divided theoretically by data selection, preprocessing, transformation, *data mining*, and interpretation.

For instance, *data mining* plays a crucial role. In this study, the KDD was done by formatting the data into a continuous time-series of the waste generation (relationship between timestamp, location and volume, in Kg), combined with socio-economic data external factors: population structure, education, activity and income, household's stage of life and income, buildings and work stages in the target area. In further analyses, a Machine Learning technique K- Nearest Neighbors (KNN) was applied to create waste generation type profiles and the clustering of areas with similar waste generations, so useful information can be extracted, from monitoring, updating waste management policies and operations in the household, real estate and utility levels.

In paper [34] it is proposed an intelligent waste material classification system, using Machine Learning algorithms, such as Support Vector Machine and Convolutional Neural Network. With this classification system, the employees could benefit from it in waste separation stages in different components with higher accuracy, which is done manually by hand-picking. When tested against their dataset, the results of accuracy reached about 87%.

# Chapter 3 - Methodology

This chapter is intended to create a roadmap on how the concepts described in the prior chapters are going to be applied in this study. To do so, the method CRISP-DM is used, which is a *data mining* methodology or process model that provides a blueprint for conducting a *data mining* project.

The first step is to understand the business objective concept, passing by:

- Assessing the situation.
- Determine *data mining* goals.
- Provide a project plan.

After the business understanding, the next step is to understand the data, starting by:

- Doing a collection of the data.
- Describe de data.
- Data exploration.
- Review the quality of the findings.

Then it is time to prepare de data, residing on:

- Select the data that fits our interest.
- Clean the data.
- Construct.
- Integrate.
- Format the data.

Modelling the data:

- Select the model.
- Test the model.
- Create the model.
- Assess the model.

Evaluation of the data:

- Evaluate the result.
- Review the process.
- Determine next steps.

### 3.1. Business Understanding

*Evox* [35] is the company that installs, configures and maintains the LoRa equipment: sensors, gateways and the server that hosts the management system. The volumetric sensors are programmed to send a measurement to the management system every time the container's door is opened.

This centralized management system provides visual information about the status of every monitored waste container. It uses a pre-defined maximum admissible residuals volume per waste container to define a collection route (see Figure 6).


*Figure 6- Current solution of the waste management system, with route optimization based on capacity available* [28].

In other words, each time the route calculation procedure is performed, it generates a new route that passes through all the containers that need trash collection. Any calculated route is therefore different from all previously calculated ones.

However, the challenge here is to predict a **uniform waste system for collection management.** Which means, a set of uniform schedules and routes that are deduced from a model of the predicted behavior of the trash containers.

In order to do that, the strategy is to identify patterns for different time periods, sent by the sensors, and correlate them with external factors, so it is possible to determine what container capacity should be installed.

### 3.2. Data Understanding

The dataset is composed of **eighteen thousand rows**, and each row records a waste volume measure, date and time, the street location, and an id of the corresponding container. In total, there are **eighteen waste containers**, identified by a unique id, its geographic coordinates, type of container and his total capacity. There are three types of containers: the standard ones, with 800 litres and 1000 litres of capacity and the surface containers, which can also store 1000 litres (see Figure 7).

*Figure 7- Standard waste containers (ids: 48843, 52910) and surface waste containers (ids: 44263, 44966, 50419).*

The containers are distributed across the district of Castelo Branco, covering about **eight streets**, as shown in Figure 8, where it shows the number of containers belonging to each street, followed by their id number and capacity. These containers are very close to each other, from up to 1 meter to 80 meters apart.


*Figure 8. Container's streets locations with a perspective view.*

The data from each container consists of the following elements:

- Container Id.
- Description.
- Container type.
- Waste type.
- Geographic localization.
- Address, localization.
- Zone.
- Date-time measures.
- Volume filled [%].

### 3.3. Data Preparation

This data must be cleaned and organized in appropriate structures to begin its mining. For this, it is decided to work with the Python, because of its simplicity to manipulate datasets.

The classification of the collection days is done according to the weather conditions and the type of day. For the weather conditions, it is used the meteorological information provided by the National Centers for Environmental Information. The temperature and rain data are divided into predefined classes.

Regarding the type of day, all events susceptible to influence the number of waste deposits was selected: weekends, holidays (New Year, Easter, Christmas…) and commemorative dates, such as Carnival, Réveillon, Father's Day.

The additional variables added are:

- Precipitation [mm].
- Air-temperature [Celsius].
- Type of day.

For the temporal variables, the following classes were created, in order to search patterns and check the evolution across the year:

- Year.
- Month.
- Day of the week.
- Hour.

The dataset containing the information of all classes provides all the necessary knowledge to study the capacity-frequency problem.

### 3.4. Data exploration

With the dataset and its variables defined, a detailed study of the contained information is then presented and the '*mining'* of the data is started, using a visualization tool.

The data is grouped by their locations mentioned before, between the dates of 08-jun-2017 to 08 jun-2018. It is possible to see that some containers do not have records in some periods of time, (see Figure 9).

Additionally, the container nº id 44263, at Praça de Camões 6, may have had a malfunction with the sensor, since it did not send any data during the Autumn season.

Despite these issues, there is still a huge amount of data to work with, and as the majority of the containers have data from the middle of autumn 2017 to the beginning of

summer 2018, it's still possible to create patterns to solve the uniform waste collection problem.

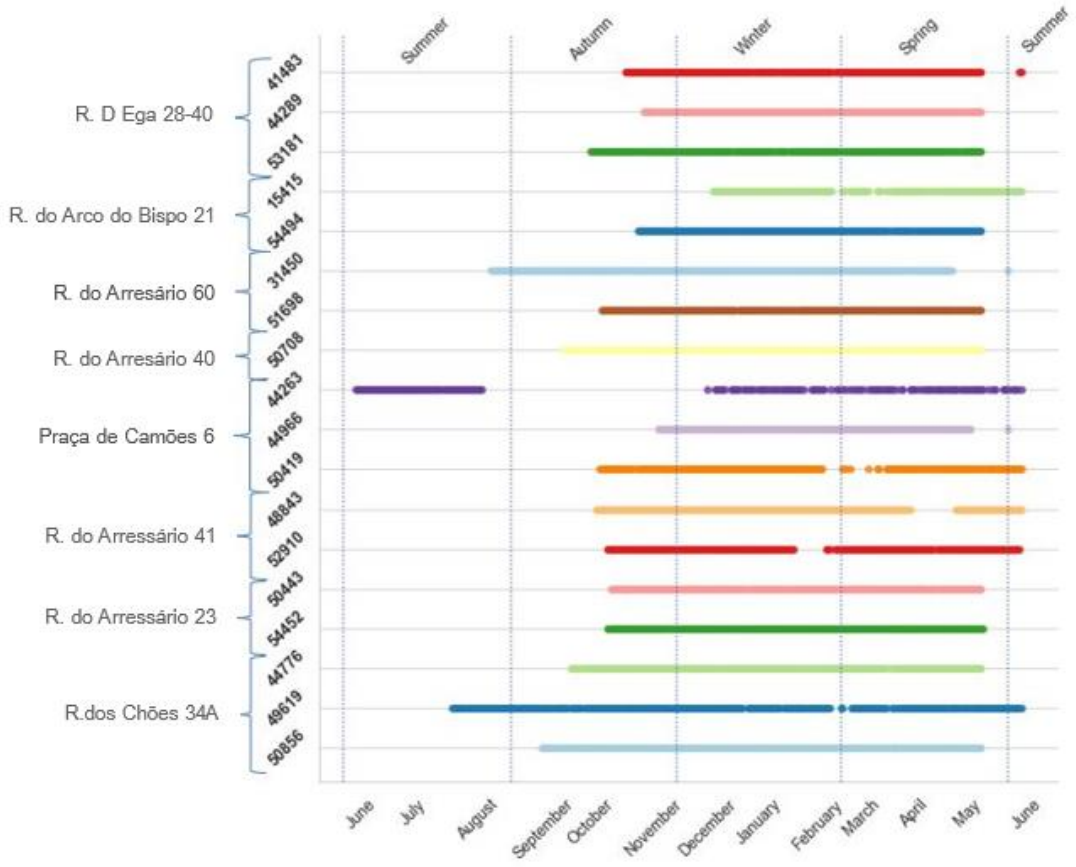


*Figure 9. Container's life cycle.*

Figure 10 shows the average volume of waste inside the containers in percentage, for each street and each month. Even for an average calculation, the values seem random. Also, their volume seems to be increasing over time, and most of the volumes are between the range of **30% to 60%.**

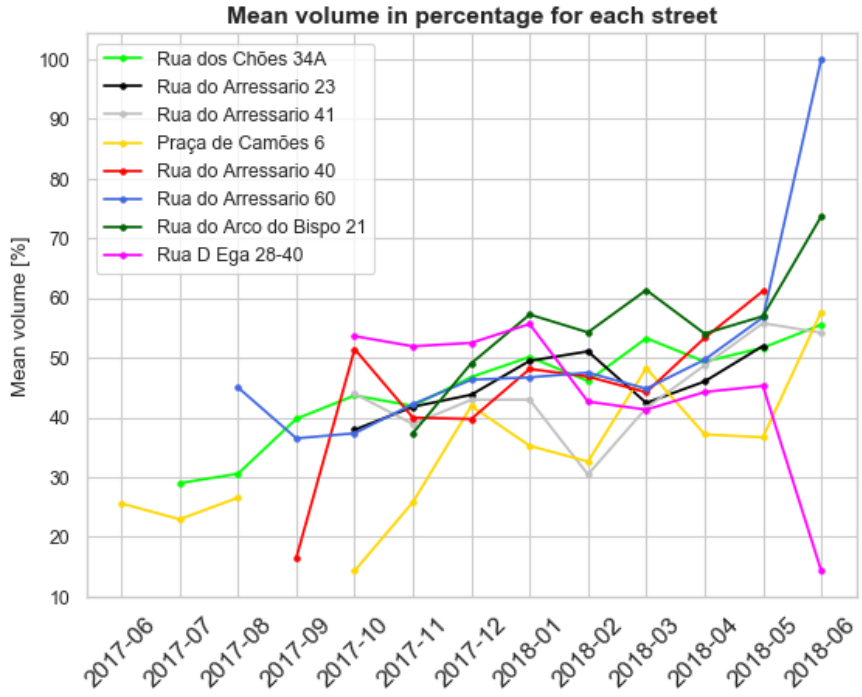


*Figure 10. The average volume of waste in the containers by street location, and their according months.*

In order to get a broader view of the data, it is necessary to check how the day of the week influences the container's volume, see (Figure 11). Surprisingly, the volume is often higher on **Wednesdays**, reaching up to **60%** volume, while for the other days, the values vary from 40% to 50% of the containers' volume.

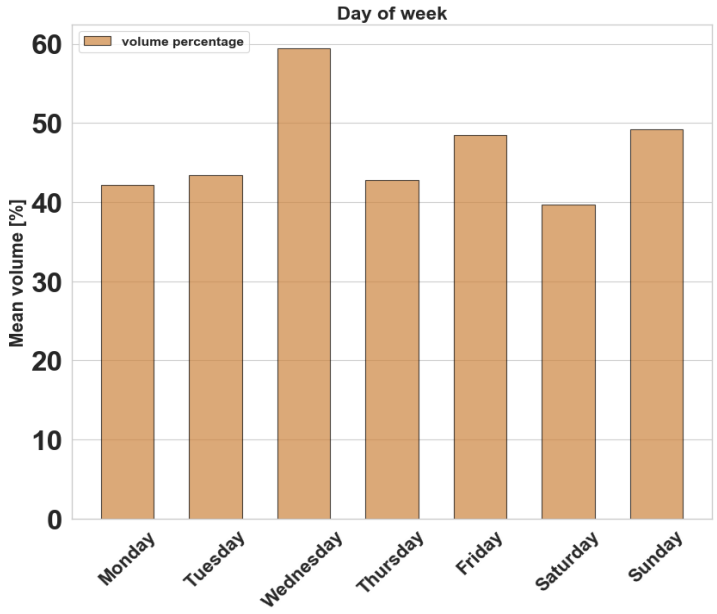


*Figure 11. Average container's volume in percentage for each day of week.*

Regarding the time of the day, (see Figure 12), the average volume of the containers [%] decreases drastically from 8h to 9h AM, which can be related to the waste **collection time**. Then, the deposits seem to be increasing linearly over time until 5 PM, increasing a bit more slowly after that.

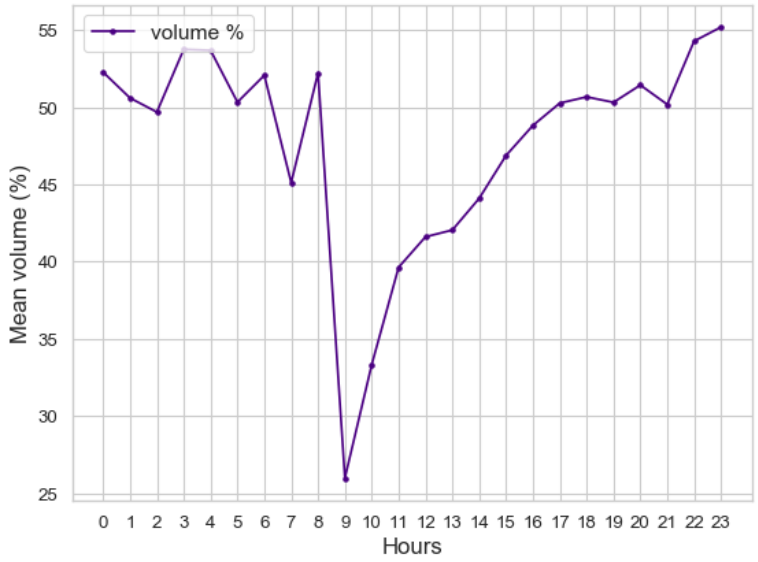


*Figure 12. Container's volume along the day.*

### 3.5. Deposits and Collections

A detailed investigation was done on to understand why the volume is higher on Wednesdays. This could be due to people depositing more residuals on Wednesdays, but this seems very unlikely. Alternatively, it could be caused by a larger gap between the two collection days.

To investigate the issue, manipulation on the dataset have been made, in order to get the **exact amount of deposits and collections**, in **liters**, for each day of week. This calculation can be expressed by the following expression:

- **Deposits**: $ɖ = i - i_{-1}, \forall i \geq 0$
- **Collections**: $ȼ = i - i_{-1}, \forall \mathrm{i} < 0$

Where:

- $i$ represents the volume of residuals, in liters, for a date-time measure.
- $ɖ$ represents the volume deposited, in liters, every time an individual deposits residual.
- $ȼ$ represents the volume collected, in liters, occurred every time the container is emptied.

After splitting the dataset in two categories, the visualization of the volumes of deposits and collections is performed. The result (see Figure 13) is the average volume (in litres) of deposits and collections, for all containers, per weekday:

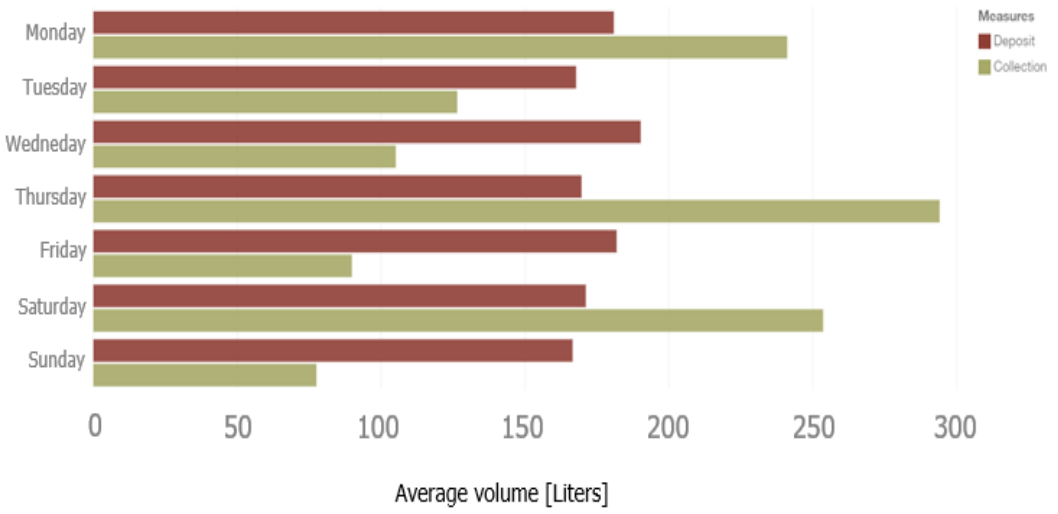


*Figure 13. The average volume of litres deposited and collected per day.*

Looking into the average volume of waste deposits, (see Figure 13), there is no notable variation between the days of the week, as they vary just from **166 to 190 litres**. That is, the volume of deposits is not influenced by the day of the week.

An interesting fact obtained by observing the collection volumes is that there are three main weekdays when the waste volume is collected - on Mondays, Thursdays, and Saturdays. From this, it has been deduced that the containers are emptied with an average frequency of **three times a week**.

Therefore, the reason why the volume of the containers, in percentage, is always higher on Wednesdays is related to the collection frequency. Which means, this is the day of the week when the interval between the two collections is higher.

For example, if a collection is done on a Monday morning (the collections are made in the morning, from 8h to 9h, according to Figure 13), then the next will only happen on the next Thursday morning, **three days** later. Conversely, from Thursday morning to Saturday or from Saturday to Monday, which is the detected collection days, there is only a gap of **two days**.

It can also be deduced that all the zones have the **same collection day programmed,** because they are very close to each other, from 40 to 80 meters.

Figure 14 shows the average volume of waste collections and deposits. Reducing the graph to the one year scale shows that the volume had two higher minimums, in August and September 2017. This may be because people go on summer vacations, in this period, leading to a lower populational density. From that time on, the deposits increase linearly until December 2017, and then the trend remains almost constant from that period (December to June).

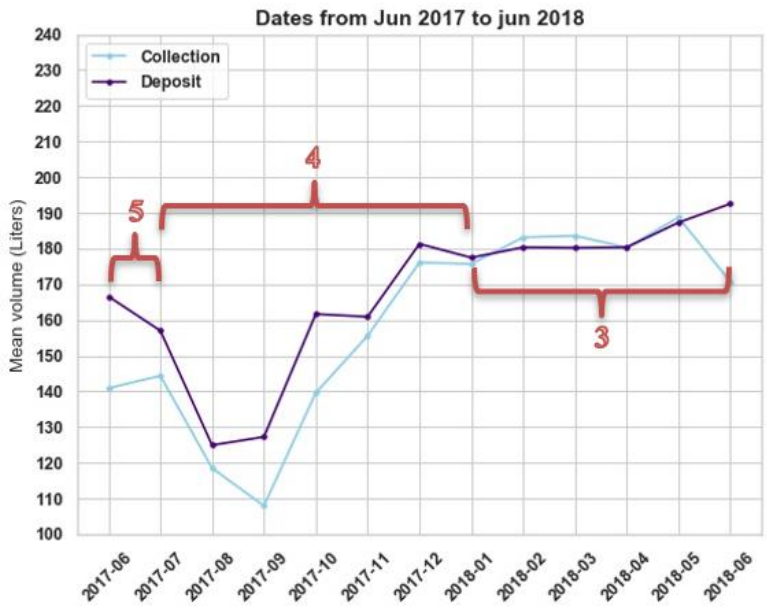


*Figure 14. The average volume of waste collection and deposit, for each month and it is waste collection frequency*

The **red marked numbers** shown in Figure 14 are the occurred **frequency per week**, for each month. It can be noticed that the frequency is dynamic, that is, it changes from month to month in order to fit the needs.

Regarding the collection frequency, it is visible that the months from July to December 2017 the frequency was four times a week: Mondays, Tuesdays, Thursdays, and Saturdays. In the months between January and June 2018, the frequency changes to three times a week since the Tuesday collection disappears.

### 3.6. Data Correlation

This section focuses on finding patterns that may influence the volume of deposits by studying **four** different scenarios, as displayed in Figure 15.

For class ***type-of-day***, the database is divided into four types of the day: the celebrative days (for example, Father's Day and *Réveillon*), the holidays, normal days, and weekends.

The class ***season*** is used to divide the database into the different seasons; class ***precipitation [mm]***, divides the database according to weather. The no rain, moderate or heavy rain occurred during each day; and class ***air-temperature [Celsius]*** divides between frosty days, cold days, warm days and hot or very hot days.

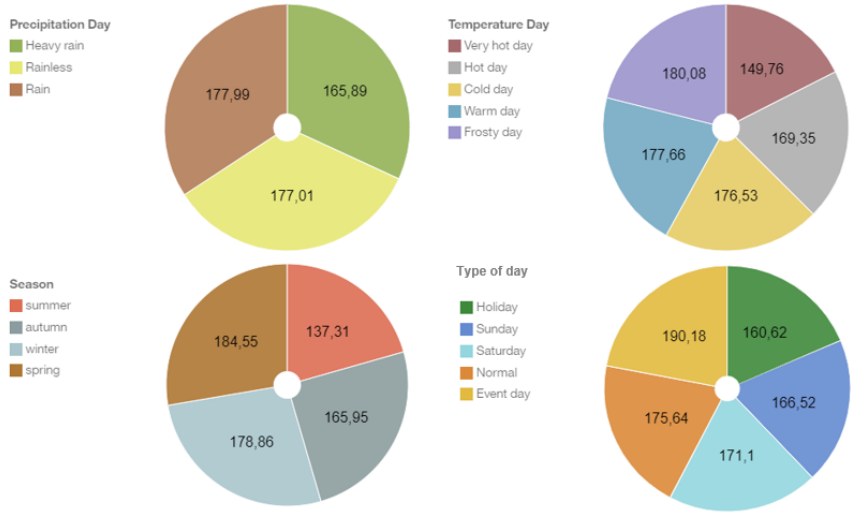


*Figure 15. Pie chart representation of the deposits by the levels of precipitation, temperature, season, and type of day.*

Regarding the levels of **precipitation**, the **average volumes of deposits** are **very close** to each other, with the lowest value of 165 with heavy rain, and the highest value of 180 with moderate rain.

Concerning the **air temperature**, the volumes of deposits range from 149 to 180 litters. The amounts are also very similar, except for those in ***very hot day***, which have a **much lower value**. This may be due to the fact that these days mostly happen during the summer holidays.

For the **season**, it can be observed that the volume of deposits in **summer** is **significantly lower** than in other seasons. As explained above, this might be because the variables *summer* and *very hot day* are correlated.

For the **type of day**, the average volumes of deposits are once again similar between the different type of days. **Their values are close to each other** and range from 160 to 190 liters.

### 3.7. Major Findings

As seen above, it is possible to extract a lot of information about the dataset with deep data visualization and analysis. These pieces of information will be used for the algorithms coming in the following sections.

Namely, it can be deduced that:

- The frequency of collection is dynamic, as it may vary according to the time of the year. Also, the days of the week for collection are often on Monday, Thursday, and Saturday. Tuesday is also added when the frequency is increased to four.
- The collections always start at 8 AM.
- The daily average volume of the containers is mostly between **30% to 60%** (330 days out of 366), and those few days where the volume is higher than 60% are almost every time on **Wednesdays**.
- The volumes of **deposits** are on average **about 180 litres per day** for each container**.** As the containers' capacities range from 800 to 1000 litters, it may be possible to decrease the frequency – especially during the Summer.
- The classes *type-of-day, precipitation [mm]* and *air-temperature [Celsius]*, **did not show good correlation results** with the volume of deposits, since the **variation was very low and quite random.**
- In the Summer period, the volume of waste deposits **decreased from 175 to 130 litters**, which may be due to less population density since families go out on holidays.
- Variable *container id* shows a good level of correlation with deposits volumes, as containers get filled more often.

- Variable *month* also shows a good correlation with deposits volumes, as far as can be seen with the available data.

# Chapter 4 - Model Optimization

## 4.1 Model Formulations for Optimizing Collection Frequency

According to the data exploration of the previous chapter, it was concluded that in most of the occurrences, the containers volume is below their average capacity (around 45%), before its collection.

So, in order to optimize the collection in a uniform system, some data manipulation needs to be performed. The specifications that are going to be defined in this thesis are:

- Determine a fixed weekly routine of collection.
- Define a fixed time-schedule for the collection.
- Consider the lower deposits in the summer season.
- Ignore collections from the non-collection days.
- Apply and exhibit the new data results.
- Increase or decrease specific container's capacity, from 800 litres to 1000 litres, or add/reduce a new container to a specific street.

To define a fixed collection-frequency for a full year, three models are proposed for a time-schedule between 7 to 9 AM:

- Model 1: for a collection of three times a week, whose collection days are on Monday, Wednesday, and Friday.
- Model 2: for a collection of three times a week, also on Monday, Wednesday, and Thursday, except on summer season: to two times a week, Monday, and Thursday.
- Model 3: for a collection of two times a week, on Monday and Thursday.

## 4.2 Comparison of the proposed models

Figure 16 which is related to model 1, shows the volume variation, (green colour) in comparison with the real volume (blue colour). As it can be observed, the new volume slightly increases on non-collection days.

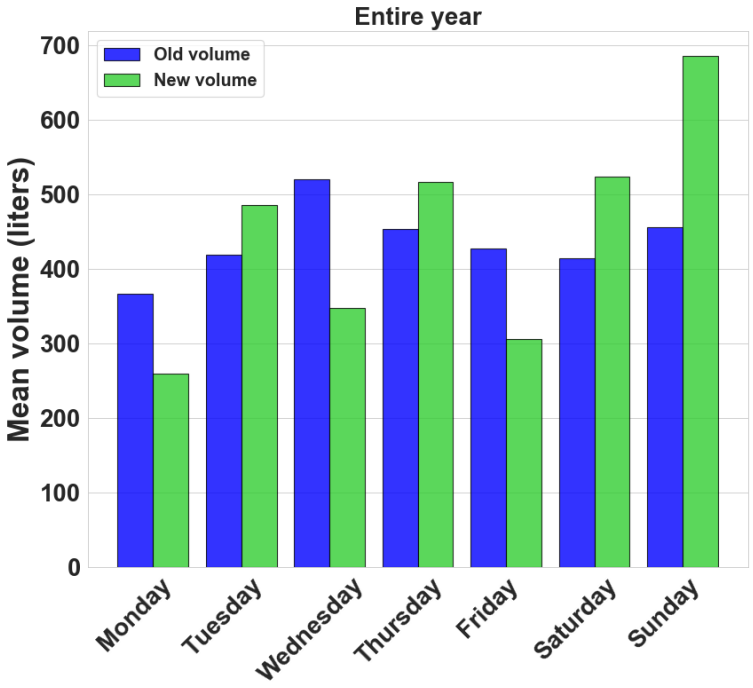


*Figure 16. Model for three times a week.*

Few changes are visible, when comparing model 2 with model 1, since they have similar formulations. The visualization is made by calculating the overall average volume of the entire year, where there is a lack of data regarding the summer season.

To avoid this issue and calculate the effective benefits of a collection of two times a week just for summer, Figure 17 shows the volume variation for these three months.

The volume reaches the highest values on weekends, however, in most of the times, this value does not exceed the maximum capacity of the container.

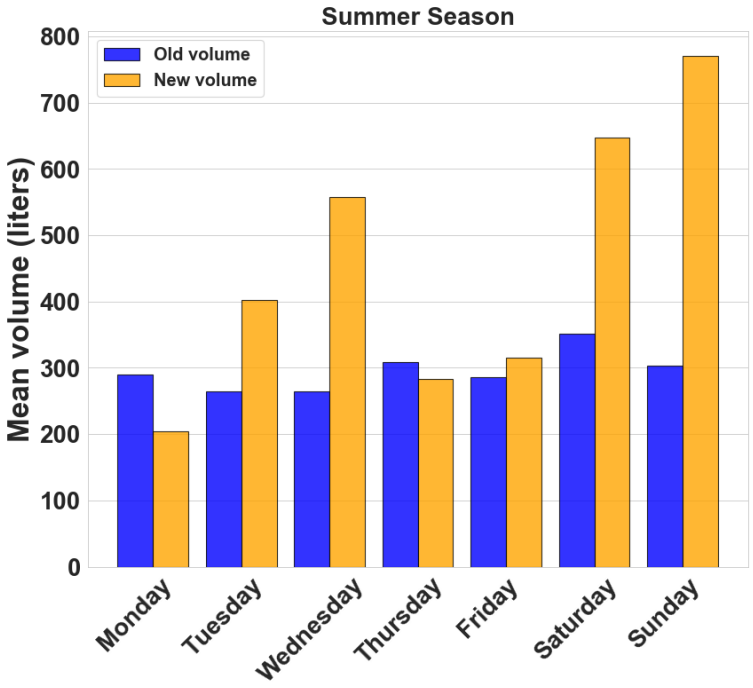


*Figure 17. Model for two times a week in summer season.*

Model 3 refers to the fixed collection of two times a week, on Monday and Thursday. From the outcomes shown in Figure 18, it can be observed that the biggest differences between the old and the new volume are on weekends.

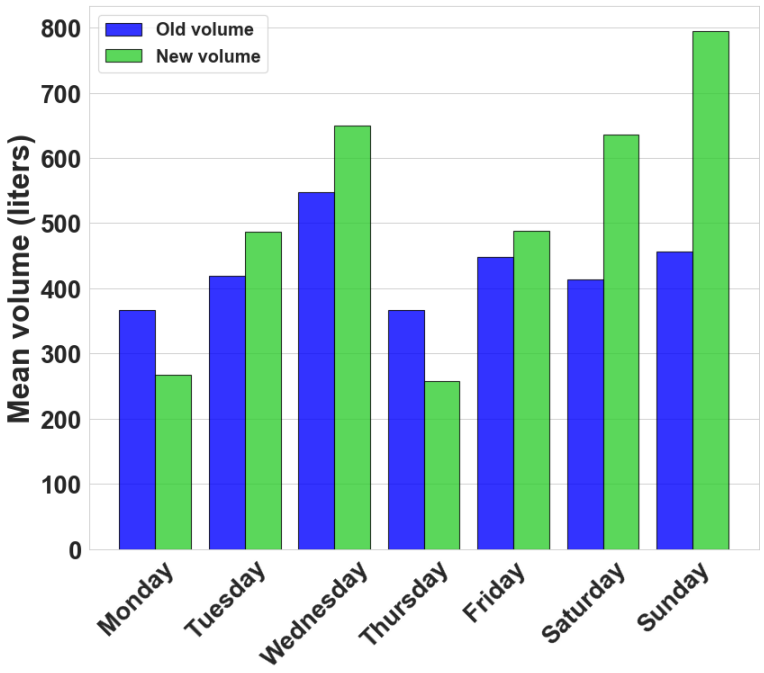


*Figure 18. Model for two times a week.*

All of the proposed models meet the requirement of not overflowing the container capacity. The decision of what is the most suitable one depends on many factors and priorities, such as sanitary requirements, environment, and economic reasons.

Taking this into account, model 2 will be used to proceed with data manipulation, because it is the one that fulfils the requirements with the minimum collection, without overflowing the containers, which in practice represents more time savings, and it's also the most environmentally friendly.

### 4.3 Application of the selected model

After selecting model 2, it is now necessary to evaluate if the number of waste containers should be increased or decreased, in order to ensure that all streets are being filled at the same speed (see Figure 19).

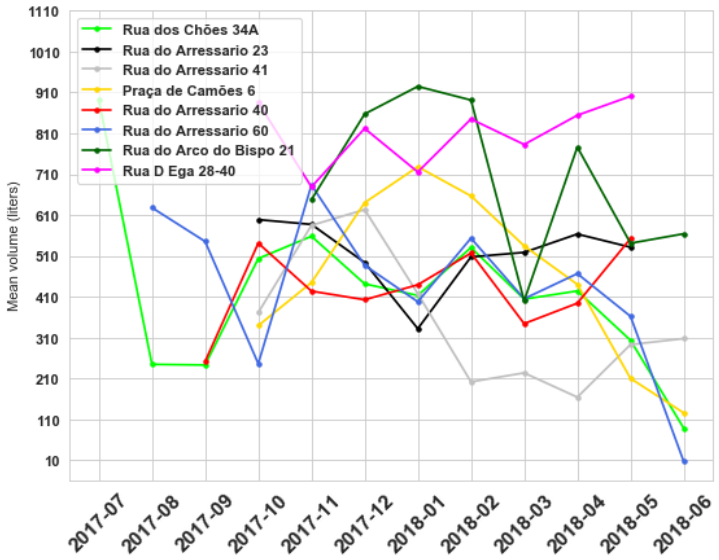


*Figure 19. Average volume for each street.*

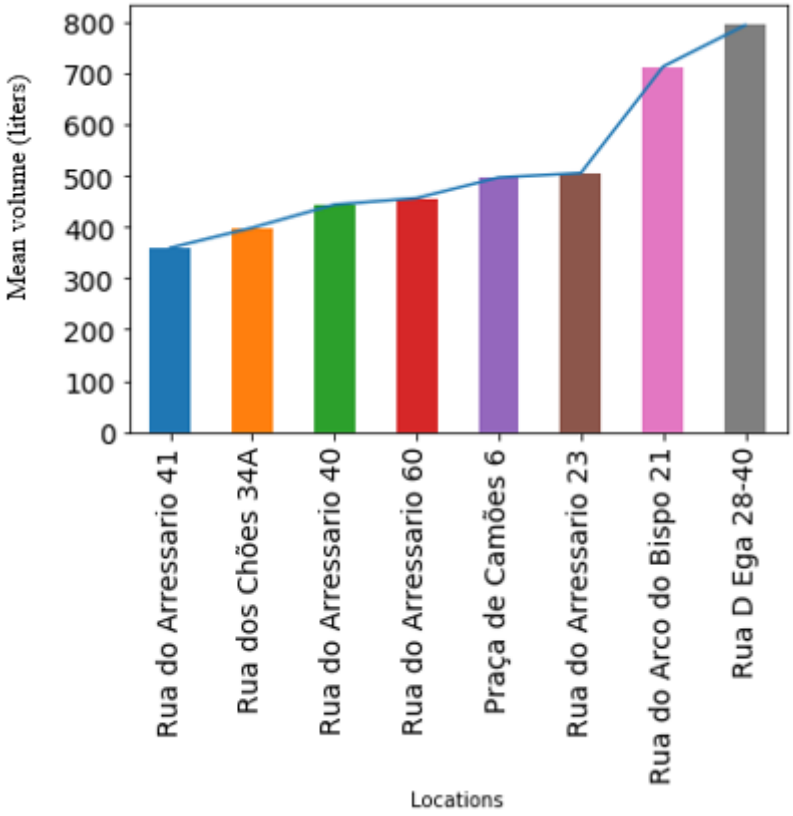


*Figure 20. Average container's volume for each street location.*

From Figures 19 and 20, it is possible to see that streets “Rua D. Ega 28-40” and “Rua do Arco do Bispo 21” have, in average, higher levels of waste volume when compared to other streets.

To proceed with the optimization, a premise must be defined - the *Street Capacity*. *Street Capacity* is the sum of the total capacity of the waste container.

- Rua D. Ega 28-40 has **three** waste containers of 1000 litres of capacity.
- Rua do Arco do Bispo 21 has **two** containers of 1000 litres of capacity.

For instance, the *Street Capacity* of these streets is 3000 and 2000 liters, correspondingly.

Several experiments were created to enhance the street capacity, considering the current average values of these two streets and the available container format, which are 800 and 1000 litres. The best results performed are displayed in Figure 21.

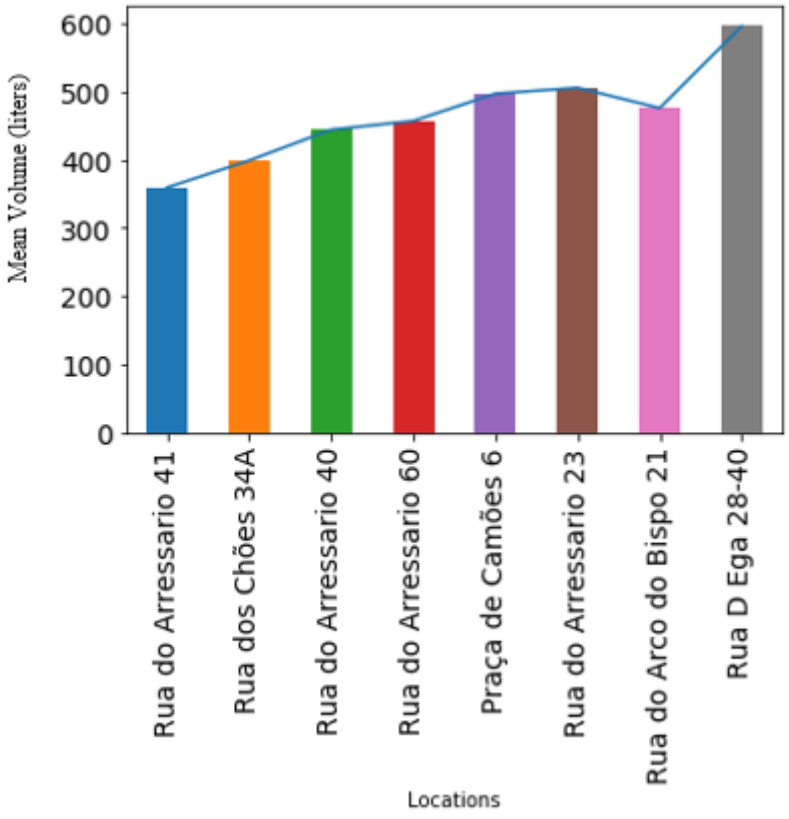


*Figure 21. Street Capacity optimization.*

In Figure 21, an extra container of 1000 litres was added to Rua D. Ega 28-40, which resulted in an improvement of the *Street Capacity* by 33.3%. The same happened for the street Rua do Arco do Bispo 21, in this case, the *Street Capacity* has an improvement of 50%.

Now the volumes are almost at the same value, around 350 to 550 litres in average. This optimization will force the streets to be filled at the same rate.

On Figure 22, the *y* axis corresponds to the value of volume in percentage. These values depend on each month and are hard to predict since there is a set of external variables and randomness, however, it is possible to filter the noise into a reasonable margin. For instance, with the addition of two waste containers, Rua D Ega 28-40 and Rua Arco do Bispo 21 are more suited to the pattern created for each month.

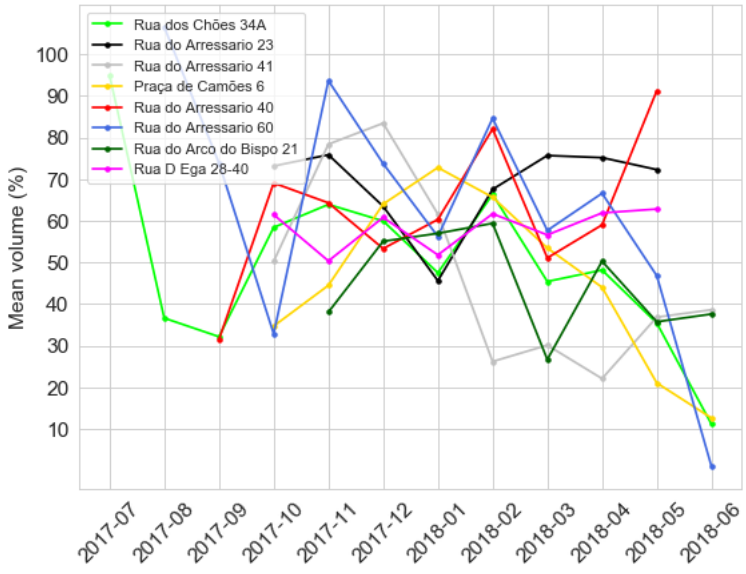


*Figure 22. Average volume for each street with optimization.*

# Chapter 5 – Prediction Process

Now that the dataset has been analyzed, the next steps towards building a prediction model are:

- Confirm that the dataset is clean and all meaningful correlations have been found.
- Define the objective that is, what variable is intended to predict.
- Determine the nature of the machine learning problem: Supervised Learning, Unsupervised learning, Semi-Supervised leaning, or Reinforced learning.
- Based on the Machine Learning problem identified, determine the type of algorithm should be used: Classification, Regression, Clustering, and others.
- Prepare the dataset so it can be used by an algorithm of the chosen type.

## 5.1. Determine the objective

The aim of the thesis is to present **capacity solutions and optimize schedules for a uniform collection system.** Accordingly, the objective is to predict when a location needs to be collected, based on the classes that were considered to be meaningful, which are: Season, Month, Street Location and the Day of Week.

From the data visualization of the previous chapter, it can be seen that most of the daily volume averages of the locations are between 30% and 70%. However, there are still cases where the volume is higher than usual (and it should).

Therefore, according to the objective, the most useful target to predict is, if a street should be collected, based on the average values of volume filled. To proceed with the application of ML, first, a question must be answered:

- What should be the threshold value for a street to be collected?

It is a subjective question, and there is not a perfect answer for it. For this case study, the limit will be defined as 60%. Figure 23 shows the number of times that this limit has been exceeded. Most of the times are at the end of the week, which is a good ML indicator.

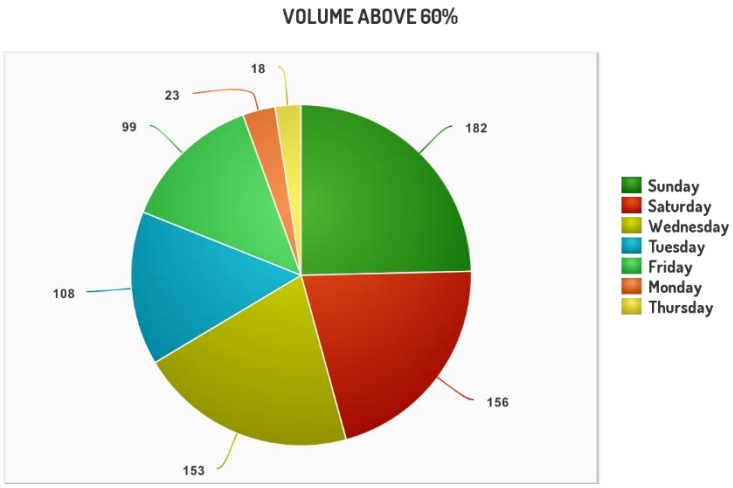


*Figure 23 - Waste volume above 60%.*

### 5.2. Choice of a learning problem and its type

From the given problem and the statements associated, this can be considered a typical **supervised** ML problem. That is, from the provided inputs - *day of week*, *season*, *street location* and *month*, the goal is to predict the output - whether the location should be collected or not (if is higher than 60% or is not), according to the expected volume for it. This matches the definition of a supervised learning problem, which consists of a set of input variables (x̧) and an algorithm to learn the mapping function from the input, in order to predict the output variable (Ŷ) [36].

$$\hat{Y} = f(x)$$

The purpose of this algorithm is to approximate the mapping function so well that new inputs of data (x̧) can be used to predict the output variable (Ŷ) for that data.

Given the learning problem defined above, the algorithm must be of **classification type**. It is not possible to determine *a priori* which one of the various *classification* algorithms better fits this problem. So, the following algorithms will be tested, and their results compared:

- Decision Trees.
- Random Forests.
- Support Vector Machine.
- k-Nearest Neighbors.

### 5.3. Data preparation

Raw data with class information, namely *street location*, *day-of-week*, *month*, and *season*, is pre-processed using *dummy* techniques, where the number of columns is equal to the number of categories.

To prepare the data to be used by the selected learning algorithms, they need to be reprocessed into records that match a set of *inputs* to an *output*. In this case, the inputs are:

- 'Street Location'.
- 'Season'.
- 'Month'.
- 'Day of week'.
- 'Volume-filled' [%].

and the output is:

- 'Needs collection'.

The output variable 'Needs collection' is a binary variable that indicates whether a street needs to be collected. To improve the performance and match the points of interest of the thesis, the volume [%], which varies from 0% to 100% is converted to binary data. When the volume filled is greater than 60%, returns 1, otherwise returns 0. This new dataset then needs to be pre-processed to align the data to the same scale, using dummy btechniques, standardization, or integer values.

Given that the hold-out method will be used to test the results, the resulting records are then randomly assigned to two different sets:

- Training data – containing 80% of the records.
- Test data – containing the remaining 20%.

The new dataset will be used to train our machine learning model through the workflow process illustrated in Figure 24.

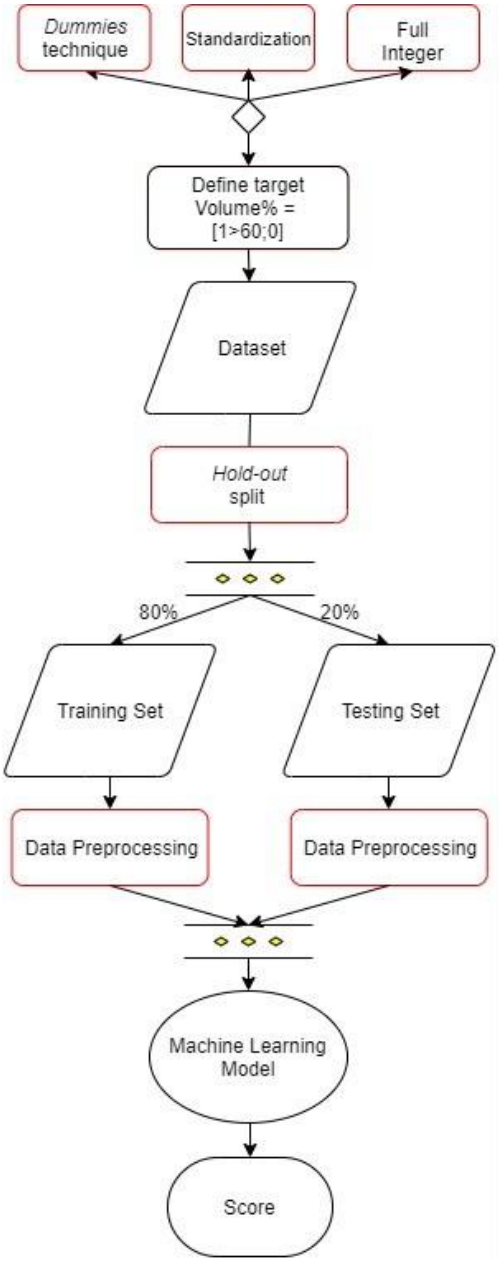


*Figure 24. Machine learning prediction process.*

### 5.4. Evaluate algorithms

In this section it is used the prepared data with the chosen algorithms, namely the *Decision Tree, Random Forest, Support Vector Machine* and *K-Nearest Neighbors* to test their accuracy in the model. The model is prepared to test if a certain location needs to be collected, based on the classes that were found important.

The metrics used to evaluate these Machine Learning algorithms are:

- Classification Accuracy.

- Confusion Matrix.
- F1 Score.

### 5.4.1 Decision Tree score

Decision Tree provides a clear way to understand the model and the behavior of the features, while keeping it simple and with high execution speed. It is calculated using the following parameters:

- *Gini* index - The degree or probability of a variable being wrongly classified when it is randomly chosen.
- Samples – The number of samples of each branch.
- Value – The number of samples that are classified as true and false.

In Figure 25 is plotted a binary tree where each node represents a portion of the data. The node that is not a leaf, splits its part of the data in two sub-parts. The root node contains all data (from the training set).

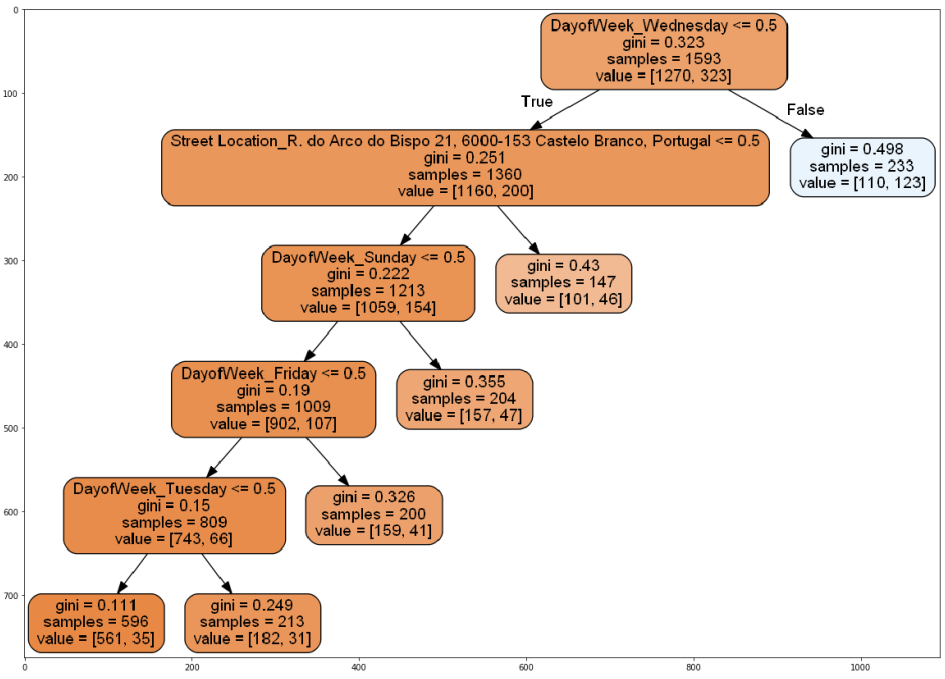


*Figure 25. Decision Tree Graph*

In this example, 1593 samples from the waste container data are predicted. The algorithm shows that the class *weekday* is the most important one, because it is shown as

the root node. This means that the day of the week, particularly on Wednesday, is when it is easier to predict the volume.

This is consistent with the findings related to the data exploration, where it was observed that Wednesdays were in average the days with more waste volume. For instance, Figure 25, shows the number of samples where the container volume is above 60%, filtered by the day of week, which is by far greater, when compared to the other days.

The second most important class is related to the street location, "R. Arco do Bispo 21", which is also an indicator that when the container volume is above 60%, the probability of been associated with this street is high.

To evaluate the model an accuracy metric has been used to calculate the precision of the algorithm (Decision Tree algorithm). The results are shown in Table 1, for the restriction metric - Number of minimum samples before each split.

*Table 1 - Decision Tree Accuracy Score.*

| Min samples split | 500 | 700 | 900 | 1000 |
|---|---|---|---|---|
| Accuracy [%] | 67.4 | 67.4 | 68.2 | 70.2 |
| F1_Score [%] | 52.9 | 44.9 | 38.6 | 40.2 |

### 5.4.2 Random Forest Score

Random Forest can be used in order to gain more scalability, that is, to reduce *variance* by generating new sets of data, resulting in a newer and bigger validation dataset. Random Forest uses the same parameters of the Decision Tree,with the addition of the following ones:*Number Of Estimators* - The number of Decision Tree classifiers.

- Criterion – How the Random Forest makes decisions. Entropy or *Gini* index.

*Table 2 - Random Forest score*

| # Estimators | 5 | 5 | 8 | 8 | 10 | 10 |
|---|---|---|---|---|---|---|
| Criterion | Entropy | *Gini* | Entropy | *Gini* | Entropy | *Gini* |
| Accuracy [%] | 71.4 | 70.9 | 70.4 | 69.1 | 69.9 | 69.1 |
| F1 Score [%] | 60.4 | 58.6 | 59.6 | 57.7 | 59.7 | 59.1 |

Similarly to Decision Tree algorithm, the Random Forest algorithm applies the same classification techniques, but in addition, multiple Decision Trees are created in order to increase the amount of data and to decrease the bias.

In this scenario, multiple tests were created. The number of estimators ranges from 5 to 10, intercalated with the decision criterion of entropy and *Gini* index.

The values of the accuracy and F1 score did not change on average. They remained between 70% and 60% respectively, which means that the bias is already big enough, so increasing the number of estimators will not have any effect.

In Figure 26 is plotted a calculation of the Random Forest confusion matrix for the experiment of 10 estimators and the criterion *Gini* index. The confusion matrix is a performance measurement of the output classes, specified for classification problems. In this matrix, four possible different combinations of the predicted values are represented:

- TP - The true positives (0,0).
- FP - The false positives (1,0).
- FN - The false negatives (0,1).
- TN - The true negatives (1,1).

All the corrected predictions are in the diagonal of the table, for instance, the TP (0,0) and TN (1,1).

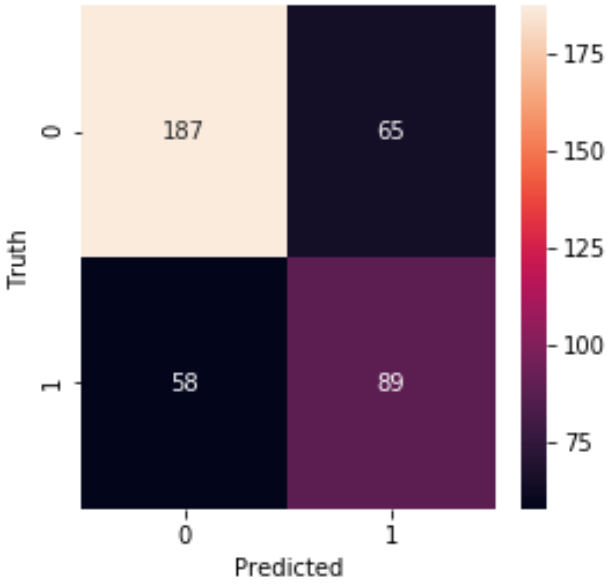


*Figure 26. Random Forest's Confusion Matrix*

With 14 estimators, the Random Forest algorithm generated 399 samples from the test size ($187 + 65 + 58 + 89$).

From those 399 samples, 189 are values that were correctly predicted, true positives (Predicted and Truth = '0'), when the container volume filled is lower than 60%.
There were 89 samples when the volume is above 60%, the true negatives. In this experiment, 123 errors were detected, which are the summary of the FP and FN ($58 + 65$).

The accuracy can also be calculated using the confusion matrix, which is the total of true predicted values divided by the total of samples, $\frac{276}{399} = 0.691$ (69%).

### 5.4.3 K-Nearest Neighbors score

Figure 27 shows the accuracy score of the *K*NN algorithm, calculated for the number of neighbours, from 1 to 30, while Figure 28 is related to the F1-Score. The highest accuracy score occurred when the number of neighbours was equal to 3, reaching a peak at 95.9% and 58% for the F1-Score.

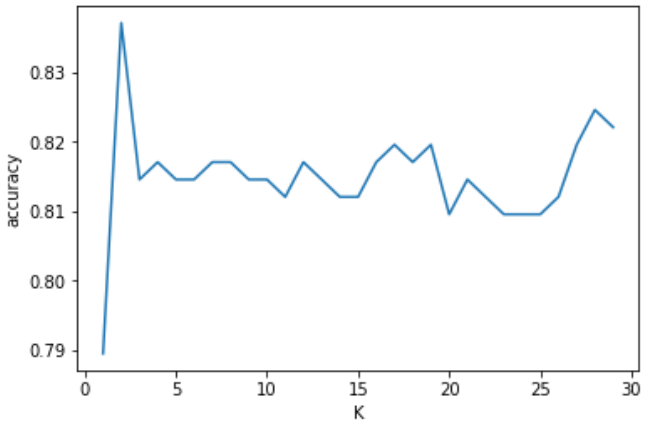


*Figure 27. K-Nearest Neighbors Accuracy Score.*

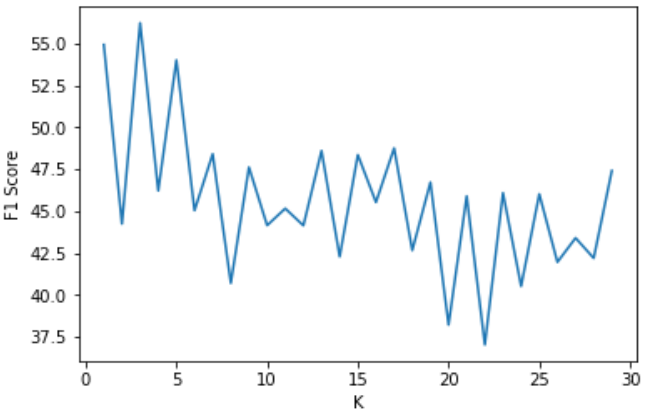


*Figure 28. K-Nearest Neighbors F1- Score.*

### 5.4.4 Support Vector Machine score

SVMs are suited for performing nonlinear classification of complex datasets with a small or medium size and work well on binary classification tasks.

The Support Vector Machine, is calculated using the following parameters:

- *Kernel* – Mathematical function to handle and transform the input data. Those functions are: Sigmoid, Radial Basis, Polynomial and Linear function.
- C – Parameter commons to all SVM kernels, trades off misclassification of training examples against the simplicity of the decision surface.
- *Gamma* - Defines how much influence a single training example has.

Using the default parameters of C = 1, and *Gamma* = 0.7, the accuracy score calculated for the chosen kernels are in Table 3:

*Table 3 – Support Vector Machine Accuracy Score.*

| Kernels | Linear | Sigmoid | Radial Basis | Polynomial |
|---|---|---|---|---|
| Accuracy [%] | 54.6 | 72.4 | 72.1 | 67.4 |

# Chapter 6 - Conclusions and Future Work

## 6.1 Conclusions

This thesis presents a waste management solution for the municipality of Castelo Branco, based on a uniform collection system. The main goal of this solution consists in the development of a system to improve the efficiency of the current waste collection, using data science and analytics, based on CRISP-DM model processing and taking into account sustainable environment factors and time savings constrains.

This system is based on a set of data records collected from the IoT sensors, installed inside each waste container. These records consist of eighteen thousand rows from a data table, containing information about the container id, a datetime, a waste volume measure, its maximum capacity, and other important variables, such as the street location. To reach an optimal solution, data analysis was performed from the records data. With the CRISP-DM method, the stages were summarized into:

- Data understanding and description.
- Data exploration.
- Data review.
- Data manipulation.
- Creation and testing the model.
- Application of ML algorithms.

To understand and describe the information, the data was cleaned, and it was created a roadmap, in order to identify the streets and to know how far they were distanced from each other. Also, to know the number of containers of each street, the distance from each container and their total capacity were considered.

In the data exploration, it was included the analysis and visualization of the data. According to the description and the objective of the problem, exploration techniques were performed using the *Python* programming language, because it handles well the data contained in a format of a table (or data frame), and it is an optimized language for executing exploration tasks, such as *group-by*, *joins*, aggregations, filtering, and other data manipulation features. For visualization of the data, the IDE Jupyter Notebook was used, since it contains a set of well-defined libraries focused on data visualization.

Additional variables were used according to the weather forecast of each day, containing the average temperature and precipitation levels of the day. The datetime was divided into year, month, day of the week and hour. Also, important dates such as holidays and celebration days were considered.

With the data visualization in a raw format (without manipulation), some conclusions were found. The frequency of collection is currently not fixed, and it varies according to the necessity of collection. The collections always start in the morning, from 7 AM to 9 AM. The average daily volume of the containers is mostly between 30% to 60%, and on Wednesday, the measurement of the container's waste volume is always higher, compared to the other days of the week. The number of daily deposits is on average about 180 liters for each container, except in the summer season, when deposits have decreased to 130 liters.

The external variables added, *type of day*, *precipitation*, and *temperature* appeared not to show any indicators or correlations regarding the measures of waste volume. An interesting factor is that some of the containers are getting filled more often, even compared with the ones in the same street.

A good indicator was the variable *month*, which showed good correlations with the volume of deposits. There were months where the population produced more residuals, and, in contrast, in summer months, the waste production was below the average.

To perform the data manipulation, the important variables were filtered according to the interest of this thesis. To optimize the current waste collection in a uniform system, multiple factors were considered:

- The number of containers of each street.
- The maximum capacity of the containers (800 or 1000 liters).
- The current daily deposits of each street.
- The deposits in summer season, which were drastically lower.
- Definition of a fixed number of collection days (from 1 to 3 days of collection).
- Definition of the days of week the collection should be performed.

Finally, the optimized model was defined for a fixed collection of two times a week, for Monday and Thursday.

The number of waste containers was increased by 1 in the streets of ‘Rua Arco do Bispo 21’ and ‘Eua D. Ega’ 28-40. The volumes with this optimization approach forced the streets to be filled at the same rate, on average 160 litres per day/container.

With the optimized model, supervised Machine Learning algorithms were performed, in order to try to predict if a certain street needs collection. The variables considered meaningful are the same ones that showed good correlations in the previous chapters, which are the *day of week*, *month*, *season*, and *street location*. It was defined that a certain street must be collected when their waste volume is above 60%, on average.

The Machine Learning algorithms used – Decision Tree, Random Forest, K-Nearest Neighbors and Support Vector Machine were applied into the model and the results were calculated according to the accuracy scores and f1-score.

The results of the accuracy score reached about 70% accuracy and 50% on f1-score, which are below than expected, especially the f1-score.

While the accuracy score is the division between the number of correct and the total predictions, the F1-Score effectuates penalties on the wrong predictions, combining the precision and the recall of the model.

The f1-score shows that the accuracy score is not robust and gives a false sense of confidence. This happens when the data is unevenly distributed, which lead to inaccurate volume levels..

During the development of the work, a paper “Optimize Capacity for a Uniform Waste Transportation Collection” was also published, in the book “Intelligent Transport Systems. From Research and Development to the Market Uptake”. INTSYS 2019 [37].

## 6.2 Future work

Although the development of the system presents optimal metrics, for the current system, it is still possible to make some improvements and combine extra information:

- **Collect more data** – The records of data were just for one year, and most of the containers do not have a full year of data. The maturity of the system will increase as more data is available to explore, leading, as well, to more reliable ML results (until a certain point).

- **Include more waste types** – This experiment relies only on data of municipal solid residuals, more information can be extracted if including records of plastic, glass, and paper measures.
- **Expand the location** – Increase the root to the entire municipality of Castelo Branco, and develop a solution taking into account the number of necessary vehicles to ensure the collection, their type, capacity and explore the viability of using hybrid vehicles, that can collect multiples types of waste.

## References

[1] "Goal 11: Sustainable cities and communities | UNDP." [Online]. Available: https://www.undp.org/content/undp/en/home/sustainable-development-goals/goal-11-sustainable-cities-and-communities.html. [Accessed: 19-Aug-2020].

[2] "68% of the world population projected to live in urban areas by 2050, says UN | UN DESA | United Nations Department of Economic and Social Affairs." [Online]. Available: https://www.un.org/development/desa/en/news/population/2018-revision-of-world-urbanization-prospects.html. [Accessed: 19-Aug-2020].

[3] B. Esmaeilian, B. Wang, K. Lewis, F. Duarte, C. Ratti, and S. Behdad, "The future of waste management in smart and sustainable cities: A review and concept paper," *Waste Manag.*, vol. 81, pp. 177–195, 2018, doi: 10.1016/j.wasman.2018.09.047.

[4] C. S. Burke, E. Salas, K. Smith-Jentsch, and M. A. Rosen, "A Global Review of Solid Waste Management," *A Glob. Rev. Solid Waste Manag.*, pp. 10–15, 2012, doi: 10.1201/9781315593173-4.

[5] "Waste - Environment - European Commission." [Online]. Available: https://ec.europa.eu/environment/waste/index.htm. [Accessed: 17-Sep-2020].

[6] N. A. Ismail, N. A. A. Majid, and S. A. Hassan, "IoT-based smart solid waste management system a systematic literature review," *Int. J. Innov. Technol. Explor. Eng.*, vol. 8, no. 8, pp. 1456–1462, 2019.

[7] L. Minh Dang, M. J. Piran, D. Han, K. Min, and H. Moon, "A survey on internet of things and cloud computing for healthcare," *Electron.*, vol. 8, no. 7, pp. 1–49, 2019, doi: 10.3390/electronics8070768.

[8] O. Eriksson *et al.*, "Municipal solid waste management from a systems perspective," *J. Clean. Prod.*, vol. 13, no. 3, pp. 241–252, 2005, doi:

10.1016/j.jclepro.2004.02.018.

[9] “Statistics | Eurostat.” [Online]. Available: https://ec.europa.eu/eurostat/databrowser/view/sdg_11_60/default/table?lang=en. [Accessed: 17-Sep-2020].

[10] M. Fleischmann, J. M. Bloemhof-Ruwaard, R. Dekker, E. Van Der Laan, J. A. E. E. Van Nunen, and L. N. Van Wassenhove, “Quantitative models for reverse logistics: A review,” *Eur. J. Oper. Res.*, vol. 103, no. 1, pp. 1–17, 1997, doi: 10.1016/S0377-2217(97)00230-0.

[11] “Reverse Logistics and its Pros & Cons – GoPigeon,” 2016. [Online]. Available: https://gopigeonofficial.wordpress.com/2016/02/22/reverse-logistics-and-its-pros-cons/. [Accessed: 24-Aug-2020].

[12] G. Laporte, “Fifty years of vehicle routing,” *Transp. Sci.*, vol. 43, no. 4, pp. 408–416, 2009, doi: 10.1287/trsc.1090.0301.

[13] “Travelling salesman problem - Wikipedia.” [Online]. Available: https://en.wikipedia.org/wiki/Travelling_salesman_problem. [Accessed: 17-Sep-2020].

[14] G. B. Dantzig and J. H. Ramser, “The Truck Dispatching Problem,” *Manage. Sci.*, vol. 6, no. 1, pp. 80–91, 1959, doi: 10.1287/mnsc.6.1.80.

[15] T. Bektaş and G. Laporte, “The Pollution-Routing Problem,” *Transp. Res. Part B Methodol.*, vol. 45, no. 8, pp. 1232–1250, 2011, doi: 10.1016/j.trb.2011.02.004.

[16] Jean-François Cordeau, G. Laporte, M. W. P. Savelsbergh, and D. Vigo, “Vehicle Routing,” no. January 2007, 2006, pp. 2–10.

[17] M. A. Hannan, M. Akhtar, R. A. Begum, H. Basri, A. Hussain, and E. Scavino, “Capacitated vehicle-routing problem model for scheduled solid waste collection and route optimization using PSO algorithm,” *Waste Manag.*, vol. 71, pp. 31–41, 2018, doi: 10.1016/j.wasman.2017.10.019.

[18] T. Henke, M. G. Speranza, and G. Wäscher, “The multi-compartment vehicle routing problem with flexible compartment sizes,” *Eur. J. Oper. Res.*, vol. 246, no. 3, pp. 730–743, 2015, doi: 10.1016/j.ejor.2015.05.020.

[19] D. V. Tung and A. Pinnoi, “Vehicle routing-scheduling for waste collection in Hanoi,” *Eur. J. Oper. Res.*, vol. 125, no. 3, pp. 449–468, 2000, doi: 10.1016/S0377-2217(99)00408-7.

[20] O. M. Johansson, “The effect of dynamic scheduling and routing in a solid waste management system,” *Waste Manag.*, vol. 26, no. 8, pp. 875–885, 2006, doi: 10.1016/j.wasman.2005.09.004.

[21] I. Markov, S. Varone, and M. Bierlaire, “Integrating a heterogeneous fixed fleet and a flexible assignment of destination depots in the waste collection VRP with intermediate facilities,” *Transp. Res. Part B Methodol.*, vol. 84, pp. 256–273, 2016, doi: 10.1016/j.trb.2015.12.004.

[22] T. Anagnostopoulos, K. Kolomvatsos, C. Anagnostopoulos, A. Zaslavsky, and S. Hadjiefthymiades, “Assessing dynamic models for high priority waste collection in smart cities,” *J. Syst. Softw.*, vol. 110, pp. 178–192, 2015, doi: 10.1016/j.jss.2015.08.049.

[23] J. Bautista, E. Fernández, and J. Pereira, “Solving an urban waste collection problem using ants heuristics,” *Comput. Oper. Res.*, vol. 35, no. 9, pp. 3020–3033, 2008, doi: 10.1016/j.cor.2007.01.029.

[24] T. R. P. Ramos and R. C. Oliveira, “Delimitation of service areas in reverse logistics networks with multiple depots,” *J. Oper. Res. Soc.*, vol. 62, no. 7, pp. 1198–1210, 2011, doi: 10.1057/jors.2010.83.

[25] A. Augustin, J. Yi, T. Clausen, and W. M. Townsley, “A study of Lora: Long range & low power networks for the internet of things,” *Sensors (Switzerland)*, vol. 16, no. 9, 2016, doi: 10.3390/s16091466.

[26] S. Dixit, “Design and implementation of a LoRa based Gateway,” 2017.

[27] "LPWAN, LoRa, LoRaWAN and the Internet of Things | by Prashant Ram | Coinmonks | Medium," 2018. [Online]. Available: https://medium.com/coinmonks/lpwan-lora-lorawan-and-the-internet-of-things-aed7d5975d5d. [Accessed: 24-Aug-2020].

[28] 360Waste, "The future of waste collection." p. 8, 2018.

[29] T. Nakahara and H. Morita, "Pattern mining in POS data using a historical tree," *Proc. - IEEE Int. Conf. Data Mining, ICDM*, pp. 570–574, 2006, doi: 10.1109/icdmw.2006.129.

[30] D. Rutqvist, D. Kleyko, and F. Blomstedt, "An automated machine learning approach for smart waste management systems," *IEEE Trans. Ind. Informatics*, vol. 16, no. 1, 2020, doi: 10.1109/TII.2019.2915572.

[31] T. Bakhshi and M. Ahmed, "IoT-Enabled Smart City Waste Management using Machine Learning Analytics," in *ICECE 2018 - 2018 2nd International Conference on Energy Conservation and Efficiency, Proceedings*, 2018, doi: 10.1109/ECE.2018.8554985.

[32] F. Ferreira, C. Avelino, I. Bentes, C. Matos, and C. A. Teixeira, "Assessment strategies for municipal selective waste collection schemes," *Waste Manag.*, vol. 59, 2017, doi: 10.1016/j.wasman.2016.10.044.

[33] H. Niska and A. Serkkola, "Data analytics approach to create waste generation profiles for waste management and collection," *Waste Manag.*, vol. 77, 2018, doi: 10.1016/j.wasman.2018.04.033.

[34] O. Adedeji and Z. Wang, "Intelligent waste classification system using deep learning convolutional neural network," in *Procedia Manufacturing*, 2019, vol. 35, doi: 10.1016/j.promfg.2019.05.086.

[35] "EVOX Technologies." [Online]. Available: https://www.evox.pt/. [Accessed: 19-Aug-2020].

[36] S. B. Kotsiantis, “Supervised Machine Learning: A Review of Classification Techniques,” *Hyperfine Interact.*, vol. 237, no. 1, pp. 1–8, 2016, doi: 10.1007/s10751-016-1232-6.

[37] J. T. Costa, A. F. Oliveira, A. L. Martins, and J. C. Ferreira, “Optimize Capacity for a Uniform Waste Transportation Collection,” in *Intelligent Transport Systems. From Research and Development to the Market Uptake*, 2020, pp. 108–128.