NOVA IMS Information Management School

MGI

Master Degree Program in

Information Management

ANALYSIS OF ILLEGAL PARKING BEHAVIOR IN LISBON

Predicting and Analyzing Illegal Parking Incidents in Lisbon's Top 10 Critical Streets

Joana Maria Gonçalves

Project Work

presented as partial requirement for obtaining the Master Degree Program in Information Management

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

ANALYSIS OF ILLEGAL PARKING BEHAVIOR IN LISBON

By

Joana Maria Gonçalves

Project Work presented as partial requirement for obtaining the Master's degree in Information Management, with a specialization in Knowledge Management and Business Intelligence.

Supervisor Professor Miguel de Castro Simões Ferreira Neto

Co-Supervisor*:* Professor João Bruno Morais de Sousa Jardim

July 2023

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledge the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Joana Maria Gonçalves

Lisbon, July 14th of 2023

ABSTRACT

Illegal parking represents a costly and pervasive problem for most cities, as it not only leads to an increase in traffic congestion and the emission of air pollutants but also compromises pedestrian, biking, and driving safety. Moreover, it obstructs the flow of emergency vehicles, delivery services, and other essential functions, posing a significant risk to public safety and impeding the efficient operation of urban services. These detrimental effects ultimately diminish the cleanliness, security, and overall attractiveness of cities, impacting the well-being of both residents and visitors alike.

Traditionally, decision-support systems utilized for addressing illegal parking have heavily relied on costly camera systems and complex video-processing algorithms to detect and monitor infractions in real time. However, the implementation of such systems is often challenging and expensive, particularly considering the diverse and dynamic road environment conditions. Alternatively, research studies focusing on spatiotemporal features for predicting parking infractions present a more efficient and cost-effective approach.

This project focuses on the development of a machine learning model to accurately predict illegal parking incidents in the ten highly critical streets of Lisbon Municipality, taking into account the hour period and whether it is a weekend or holiday. A comprehensive evaluation of various machine learning algorithms was conducted, and the k-nearest neighbors (KNN) algorithm emerged as the top-performing model. The KNN model exhibited robust predictive capabilities, effectively estimating the occurrence of illegal parking in the most critical streets, and together with the creation of an interactive and user-friendly dashboard, this project contributes valuable insights for urban planners, policymakers, and law enforcement agencies, empowering them to enhance public safety and security through informed decision-making.

KEYWORDS

Smart Cities; Illegal Parking; Urban Transportation; Urban Planning; Predictive Modeling; Decision-Support

Sustainable Development Goals (SGD):

INDEX

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

LA	Lisboa Aberta
CML	Camara Municipal de Lisboa
WHO	World Health Organization
CRISP-DM	Cross Industry Standard Process for Data Mining
MSE	Mean Squared Error
RMSE	Root Mean Squared Error
MSE	R-squared
DT	Decision Trees
RF	Random Forest
GB	Gradient Boosting
RFE	Recursive Feature Elimination
NN	Neural Networks
KNN	K Nearest Neighbors
LN	Linear Regression
LSS	Lasso
RID	Ridge
ELAS	Elastic Net
BAY	Bayesian
AB	AdaBoost
BG	Bagging

1. INTRODUCTION

1.1. CONTEXT

Over the past few centuries and particularly in recent decades, there has been a mass migration of populations from rural to urban areas. In 2021, urban areas were already home to 56 per cent of the world's population (*World Cities Report 2022*, 2022), and that figure is expected to grow to 68 per cent by 2050 (Ritchie & Roser, 2018). This rapid urbanization is intertwined with several existential global challenges thus effective planning, management, and financing are critical to mitigate them and develop sustainable, secure, and healthy cities (Vlahov, 2002).

In this manner, mobility problems linked with an excess of private cars are one of those crucial challenges that come into prominence in cities (Lin & Du, 2015). Since the number of vehicles increases with the huge influx of the urban population, the disparity between the rapid increment of those vehicles and limited new parking facilities results in huge parking difficulties (Liu et al., 2012).

Additionally, as the rate of people owning their vehicles increases, finding a parking space has become one of the major pain points for citizens (Jennath et al., 2019). Sixteen studies conducted between 1927 and 2001 have indicated that a majority of drivers spend between 3.5 to 14 minutes in a typical search, increasing overall traffic levels in the areas studied by 8% to 74% (Shoup, 2006) bringing a high degree of frustration and anxiety to the individual drivers and having a detrimental impact on the efficiency of the whole transportation system (Weinberger et al., 2020).

Moreover, this increase in private vehicles together with the limited parking availability also leads to an increase in traffic congestion, fuel waste, environmental damage (Basu & Ferreira, 2020; Jennath et al., 2019; Kotb et al., 2017), and can also lead drivers not to park in legally designated parking spaces on the streets (Basri Said & Syafey, 2021; Spiliopoulou & Antoniou, 2012).

Rising incidents of illegal parking have led to major increases in negative externalities in the cities associated with transportation such as traffic congestion and air pollution (Basu & Ferreira, 2020; Liu et al., 2012). Illegal parking can cause traffic congestion in several ways (Bahrami et al., 2021). For instance, these illegalities can block traffic lanes, making it difficult for other vehicles to pass through causing traffic to back up and slow down, mainly on narrow or busy streets (Weinberger et al., 2020).

Consequently, high levels of traffic congestion contribute to an increase in travel times (Fulman et al., 2020; Zoika et al., 2021), fuel consumption and emissions of pollutants such as carbon monoxide (Kotb et al., 2017; Zoika et al., 2021), and delays in responding to ambulances, fire trucks, and police cars emergencies (Nourinejad et al., 2020). Additionally, it can also make it difficult for goods and services to be transported efficiently, which can slow down commerce and reduce the economic productivity of these areas (Harriet & Poku, 2013; Weisbrod et al., 2003). All these effects linked with the increased drivers' stress and frustration (Hennessy & Wiesenthal, 1999), and city noise levels (Kumar et al., 2014) caused by the traffic congestion can negatively impact the overall quality of life of citizens.

Moreover, illegal parking can affect accessibility in the cities by drastically reduce road resources in the form of road width and capacity (Parmar et al., 2020), leading to an inconveniency of people walking on the road (Kotb et al., 2017). This can occur when parked cars block traffic lanes, bike lanes, or

sidewalks, or when parked cars make it difficult for other vehicles to maneuver. Thus, the appearance of urban areas will be degraded making them less attractive to tourists and citizens (Marsden, 2006).

Overall, Illegal parking can have significant negative impacts on tourism, traffic flow, life quality, safety, and emergency response time, therefore it is crucial to develop and implement proper parking strategies and regulations to mitigate them and consequently the referred problems.

1.2. RESEARCH GAP

To address the problem of illegal parking, numerous studies have developed decision-making systems to monitor or predict these infractions and provide solutions for managing them, as well as support policy design and implementation. Many of these studies rely on cameras and video-processing algorithms for real-time monitoring (Chen & Yeo, 2019), but this method faces challenges such as weather conditions, obstructions, lack of light, and high hardware costs, as well as being limited in its spatial coverage. As an alternative to video-based systems, other studies have proposed a more cost-efficient, data-driven approach that considers various factors contributing to illegal parking, such as temporal and weather conditions, street characteristics, and proximity to points of interest, using them as inputs for predictive models (J. Gao & Ozbay, 2017; S. Gao et al., 2019).

Therefore and in a very similar way to the study of Jardim et al. (2022) that differentiates between the different types of illegal parking and it is applied in Lisbon, this project proposes a data-driven framework to understand and predict the spatiotemporal legality of parking in the top 10 most critical streets in Lisbon city, considering the hour period and the type of day, whether it is a weekend or a holiday. In addition, this study also pretends to include other relevant explanatory variables, such as air pollution and to test additional types of feature selection and predictive models attempting to achieve more reliable forecasts.

1.3. STUDY OBJECTIVES

The motivation for the development of this work project is related to the growing need for the major cities in the world and their citizens to deal with the increase in illegal parking that leads to various negative impacts. As the capital of Portugal, Lisbon is particularly affected by this problem, due to its rapid urbanization and high demand for parking in areas with limited availability.

In this manner, the main goal of this study is to uncover the patterns and impacts of illegal parking in the city of Lisbon by spatial unit and time of the day, while trying to provide authorities with an insight to support decision making regarding parking surveillance.

In order to achieve the overarching goal, the following specific objectives have been established:

- Create a predictive model capable of forecasting the number of illegal parking incidents in the top 10 most critical roads of Lisbon by hour period and type of day, with the aim of anticipating and preventing potential parking violations in the future.
- Develop a Business Intelligence tool that features a user-friendly and interactive dashboard that provides valuable insights into the illegal parking behavior in Lisbon and assists in the decision-making process.

1.4. Study Relevance and Importance

As aforementioned, illegal parking can have several negative externalities. On quality of life of citizens matters, it namely affects their safety, mobility, and accessibility. On operationalities matters, its occurrences can (1) block emergency vehicles, (2) limit the flow of traffic, and on environmental matters, it can contribute to air and noise pollution, and negatively affect the aesthetics of the cities.

As a consequence, by studying the patterns of illegal parking, through the determination of the causes of these infractions, and being able to predict them for the top 10 most critical roads in Lisbon, this study can contribute to local authorities to respond to unforeseen situations more promptly accordingly to the street, hour period and type of day.

Therefore, this research enables responsible authorities to implement more precise and successful parking management policies in Lisbon, such as implementing dynamic parking regulation which adjusts parking prices based on period of the day and location, to influence driver behavior and reduce congestion. Additionally, the city can run awareness campaigns to educate drivers on parking regulations and its negative impacts. This also allows for the optimization of police deployment for street patrolling in real-time, and the efficient allocation of resources per hour period, while promoting sustainable transportation and developing suitable parking strategies and regulations.

In summary, predicting illegal parking can help prevent the negative impacts associated with it in the areas that are more challenging and improve Lisbon authorities' daily operations.

1.5. Research Methodology

The methodology chosen for the development of this work project is called Cross-Industry Standard Process for Data Mining (CRISP-DM), which is widely recognized and used as a guiding framework for data analysis and modeling process in the field of data mining and analytics. This process provides a systematic and structured approach to effectively address our research questions, derive meaningful insights, and contribute to the existing knowledge in our domain. According to Chapman et al. (2000), the CRISP-DM process encompasses six distinct phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment, which will be further explained in detail in the Methodology chapter.

1.6. Project Structure

To conduct this study in a more structured and coherent way, it was organized into six chapters. Initially, chapter 1 consists of the presentation and explanation of the research topic, its importance, and contributions, as well as the main objectives of this project. In chapter 2 the review of the recent literature regarding illegal parking behavior and prediction is presented. In chapter 3 the methodology used is stated, its applicability is justified according to the scope of the project, and the data collected and employed is presented and described. Chapter 4 involves the presentation of the facts of the study such as the description of how the results look like and the steps on how the data was collected and treated. To conclude, chapter 5 involves the interpretation and discussion of the findings by relating them to the research aims, questions, and objectives, and the last section, chapter 6 consists of the conclusions taken from this study, and the presentation of some future suggestions.

2. LITERATURE REVIEW

This chapter intends to give a short introduction on broader topics such as smart mobility, and parking issues using previous studies that are highly trendy. After this short context, a deep dive into the existing literature is provided for the main topic of this study.

2.1. SMART CITIES

A smart sustainable city leverages information and communication technologies, among other tools, to enhance quality of life, optimize city operations and services, and increase competitiveness, all while ensuring that it meets the needs of present and future generations (United Nations, 2019). Its components include various areas such as transportation and mobility, education, healthcare, public administration, security, infrastructure, among others (Šiurytė & Davidavičienė, 2016). Indeed, Portugal has already started to implement some of these strategies to uplift this leaver of growth for the country ("National Smart City Strategy," 2022).

Over the past few years, this has become a hot topic, not only in Portugal, but also on recent high-governance discussions on their agendas. As a matter of example, the United Nations Organization established in 2015 an Agenda for the year 2030, on a global scale, formed by 17 Goals for Sustainable Development from which the goal 11 of the Agenda is directed to the theme 'Sustainable cities and communities' (*Transforming Our World*, 2015) Therefore, policymaking in smart mobility within urban areas has now embraced the notion of considering transportation on par with climate and energy policies

2.2. URBANIZATION AND PARKING PROBLEMS

With urbanization to become one of the mega trends for the next years, several issues, from overcrowding to air pollution, extra stress on natural resources and loss of habitats to grow more food need to be attained to guarantee a smooth transition to the new normal through prevention and timely measures, and here, the Literature should play an important role.

One of these issues is parking, which means that parking slots' supply may be squeezed by demand as the growth of motor vehicle ownership throughout the world occurs. Well, on the bright side, registrations of commercial vehicles tend to follow GDP growth more closely, given that demand for road transport is directly related to the state of the economy, i.e., it is a leading indicator of economic growth ("Vehicle Sales Mirror Economic Growth (2008-2021 Trend)," 2020). However, this has negative externalities, leading to various traffic problems, solutions to mitigate issues of traffic safety, congestion, noise, air pollution, and parking, are becoming increasingly urgent (Nguyen et al., 2018).

Looking to the Portuguese case, according to IMT, in 2018, Portugal, increased the number of motorized vehicles by 4% with 6.7 million vehicles and had an 5.8% growth in the number of vehicles registered (*IMT - Manutenção*, 2018). Moreover, regarding the consequent increase of traffic congestion related to the increase of private car, in 2019, Portuguese drivers spend a daily average of 42 min in urban traffic and an average of 160 h each year in traffic jams (Group, 2019).

Therefore, this is an issue that needs to be tackled from scratch. There are plenty of studies that provide alternatives for solving this parking challenge. One example is a Mckinsey's report on mobility that suggests using smart parking-technology networks, that were successfully tested in San Francisco, once these networks connect vehicles to infrastructure and inform users where parking is available, reducing the amount of time needed to find a space. (*The Road to Seamless Mobility | McKinsey*, 2019)

2.3. ILLEGAL PARKING

As a consequence of the aforementioned rapid growth of private vehicles ownership jointly with the limited imbalanced supply of parking slots, citizens tend to have difficulties in finding available parking slots, ending up parking their cars in unauthorized or illegal places nearby their point of destination, instead of searching for legal (often costly) parking space (Zoika et al., 2021).

These illegal occurrences not only increase traffic congestion but also drivers' frustration, safety hazards, fuel waste and air pollution. All this combined creates a major impact on the livelihood quality of citizens as well as affecting directly the sustainable development efforts of cities.

Accordingly with Morillo & Campos (2014), on-street illegal parking reduction should be one basic pillar of mobility policies, which would permit a higher road capacity and, therefore, greater traffic fluidity. Moreover, Galatioto & Bell (2007) found in their examination of Palermo that illegal parking causes a rise in congestion from 50% to 200%. They also point out that road links with a high incidence of illegal parking experience more congestion, leading to a heightened production of CO2 emissions.

Kladeftiras & Antoniou (2013) study has concluded that limiting double-parking in Athens could result in an increase in speeds of about 10% to 15% and a decrease of about 15% and 20% in delay and stopped time, respectively. Additionally, they also estimated that by eliminating illegal parking in Athens, the improvements would be even greater, i.e., the average speed would increase by up to 44%, while delay and stopped time would decrease by up to 33% and 47%, respectively.

In New York, Gao & Ozbay (2016) indicated that hourly travel time increases by 3.1%, 13.6%, 20.5%, and 27.6% respectively with every increase of 1, 2, 3, and 4 vehicles double-parked per 15 minutes. Illegal parking is a significant contributor to congestion, causing 47 million hours of delay annually in the United States (Nourinejad et al., 2020). Such parking violations obstruct travel lanes and pose hazards to pedestrians and cyclists. In New York, conflicts between trucks and cyclists due to illegal truck parking average at 14% on city streets (Conway et al., 2013). Moreover, illegally parked vehicles often obstruct fire hydrants, hindering emergency response operations (Nourinejad et al., 2020).

In the bid to reduce parking illegalities and alleviate these challenges, information and communication technologies, big data and data science emerge as necessary tools for the urban metropolitans. These advanced technologies can be applied to measure and track many aspects of urban life, such as transportation and the environment (Jardim, Castro Neto, et al., 2022), while being used to develop and implement effective policies through the use of decision-making systems.

The literature has been addressing these decision-making systems in the following two distinct approaches, one relying in video-based methods and another more focus on data driven techniques.

2.3.1. Video-Based Approaches

A major amount of effort has been devoted to tackle illegally parked vehicles using video surveillance.

Following the release of the Imagery Library for Intelligent Detection Systems, Boragno et al. (2007) presented a commercial solution implemented by Ipsotek Limited to automatically recognize behaviors in a scene in real-time and, in this way, to detect a parked vehicle in a prohibited parking zone sixty seconds after it has become stationary. Additionally, in that year, Porikli (2007) introduced a reliable and computationally efficient method for detecting abandoned objects and illegally parked vehicles using public datasets. The background model in their study resembles adaptive mixture models, but instead of a combination of Gaussian distributions, each pixel is defined as multiple 3D multivariate Gaussians.

Most recently, some works have focused on the use of deep neural networks to identify illegal parking. Ng et al. (2018) proposed the implementation of iConvPark, with the use of Convolutional Neural Network as the classifier, to automatize the detection of illegally parked vehicles by providing real-time notification regarding the occurrences and locations of illegal parking cases, based on live parking lot image retrieved via an IP camera.

Similarly, other advanced video-processing techniques have been employed to illegal parking. Chen & Yeo (2019) proposed a framework, that comprises object detection and movement tracking, for automatic detection of illegally parked vehicle. More specifically, this study adopts the object detection algorithm You Only Look Once (YOLO) to detect vehicles and template matching methods using normalized cross correlation for movement tracking.

Recent works have also used video-processing methods for license plate recognition like Yin et al. (2019) that provided a method to estimate vehicle parking locations using videos of security patrolling captured in real world parking lots. End-users can define restricted zones via a map-based interface and all vehicles located in these areas can be efficiently identified once patrolling videos are received.

Although these techniques have proven usefulness and success, their implementation faces difficulties due to road conditions such as weather changes, obstructions, and lighting, and their applicability have limited coverage and can only track a limited area. Additionally, many local authorities lack the resources and infrastructure necessary to implement vehicle tracking systems and process the data. Moreover, the emphasis on real-time tracking in these studies means that authorities may not have information on the frequency of illegal activities, or the risks posed by each street until they receive the camera data and analyze it using the video-processing model.

Table 1 – Summary of the Previous Video-Based Studies

Research Objective	Approach	Reference
Design an effective commercial software for CCTV video surveillance	Implement Ipsotek's Visual Intelligence Platform for behavior recognition and alarm detection	(Boragno et al., 2007)
Develop a robust abandoned object detection method	Utilize Bayesian update mechanism with long- and short-term backgrounds for accurate segmentation	(Porikli, 2007)
Develop a vision-based system for real-time detection of outdoor illegal parking	Use Convolutional Neural Network on Raspberry Pi with IP camera for vehicle identification	(Ng et al., 2018)
Develop an automatic surveillance system for detecting unauthorized parking	Proposed framework uses object detection through You Only Look Once (YOLO) and movement tracking through template matching	(Chen & Yeo, 2019)
Develop a web-based platform for finding of illegally parked cars	Combine GIS and computer vision to extract license plate numbers and estimate parking locations	(Yin et al., 2019)

2.3.2. Data-Based Approaches

To overcome the limitations of video-based systems, researchers have developed techniques that incorporate both spatial and temporal information to predict illegal parking and assess the risk of such infractions. These spatiotemporal feature-based methods have been utilized to improve the precision and efficiency of detecting and preventing illegal parking.

To predict parking availability, Zheng et al. (2015) study used parking availability data collected from two major cities, Melbourne and San Francisco, implemented three algorithms (regression tree, neural network and support vector regression) on the two city datasets and compare the performance of those models. Also, in 2017, Google Artificial Intelligence research team used a unique combination of crowdsourcing and machine learning to build a system that can provide the driver with parking difficulty information for his destination. They created a logistic regression Machine Learning model and utilized anonymous aggregated trajectory data from mobile users who opt to share their location data. Using this, Google launched a new feature for the Google Maps App across 25 US cities that offers projections about parking difficulty close to users' destination (*Using Machine Learning to Predict Parking Difficulty*, 2017).

In the same year, J. Gao & Ozbay (2017) introduced a novel data-driven framework for understanding the influential factors and estimating the actual frequency of double parking in New York City. The study applied three feature selection methods, LASSO, stability selection and Random Forests techniques and found that the top five factors influencing double parking incidents are the quantity of hotel rooms, hourly traffic volume, size of the surrounding commercial district, length of the block, and availability of curbside parking spots. Moreover, S. Gao et al. (2019) used different machine learning algorithms such as Multiple linear Regression, Support Vector machines Decision Tree, Random Forest, Gradient Boost RegressionTrees and Deep Neural Network to predict the legitimacy of on-street parking spaces at a given time and date and their locations. They resorted to 10.8 million parking violation tickets from New York and gathered temporal features, points of interest in the region, features related to the characteristics of the area, and features of human mobility from user's smartphone GPS history records.

Still in the City of New York, Jiang et al. (2020) proposed a novel deep learning framework, called Attention-Based 2-layer Bi-ConvLSTM model to predict the number of illegal parking events in urban spaces. They collected 685426 records of illegal parking events from 2015 to 2019 and used several features such as hourly weather, traffic volumes, road network and points of interests. Additionally, in a different country, Zoika et al. (2021) investigated causes leading to illegal parking in Greek cities, with a focus on space availability, road geometry and the balance between parking demand and supply. In this way, they used Google Street View images as the main data source and developed multiple linear regression models investigating factors explaining illegal parking density.

Furthermore, Jardim et al. (2022) develop an Illegal Parking Score to measure the risk of illegal parking in Lisbon based on spatiotemporal conditions and taking into consideration the different types of parking illegalities. Therefore, similarly with the previous study, this project also considers the different types of illegalities but analyzing particularly the 10 most critical roads in Lisbon by hour period and type of day, adding additional exploratory variables to the predictive models like air pollution data and adding additional feature selection techniques and machine learning algorithms.

Overall, the goal of this work project is to offer Lisbon authorities a tool to assist in parking surveillance decision-making and to aid in the implementation and assessment of parking regulations to improve the quality of life of its citizens across Lisbon but principally, in the most problematic areas.

Table 2 – Summary of the Previous Data-Driven Studies

Research Objective	**Approach**	**Reference**
Understand and predict parking occupancy rate in smart cities	Analyzing real-time parking data from San Francisco and Melbourne using regression tree, SVR, and neural network models	(Zheng et al., 2015)
Develop a parking difficulty prediction system to assist users in finding parking spaces	Gathered high-quality ground truth data through crowdsourcing, used anonymous aggregated data, and developed robust features for training a logistic regression model	(Using Machine Learning to Predict Parking Difficulty, 2017)
Understand double parking in urban areas	Utilize parking abuse tickets, service requests, social media data, and street characteristics. Apply LASSO, stability selection, and Random Forests for feature selection and prediction	(J. Gao & Ozbay, 2017)
Understand and predict on-street parking legality in metropolitan areas	Develop a data-driven framework using NYC parking tickets data, POI data, and human mobility data. Apply Random Forest model, for prediction and classification tasks	(S. Gao et al., 2019)
Predict the number. of illegal-parking events in urban spaces	Propose Att-2BiConvLSTM model, incorporating dynamic training and attention mechanism	(Jiang et al., 2020)
Investigate causes of illegal parking and propose mitigation measures	Utilize Google Street View images as data source and employ multiple linear regression models to analyze factors influencing illegal parking density	(Zoika et al., 2021)
Develop a IPS to measure the risk of illegal parking based on spatiotemporal conditions	Use spatiotemporal features and Light Gradient Boosting Machine model to calculate IPS	(Jardim, Alpalhão, et al., 2022)

3. METHODOLOGY

3.1. CROSS INDUSTRY STANDARD PROCESS FOR DATA MINING

As previously stated in the introduction, the chosen methodology is the Cross Industry standard process for Data Mining. This model is an open, industry-standard methodology that provides a structured and systematic approach to the entire data mining process. It consists of several distinct phases that take a project from its initial conception to its final deployment, and it is designed to be flexible and adaptable, allowing organizations to tailor the methodology to their specific needs and objectives.

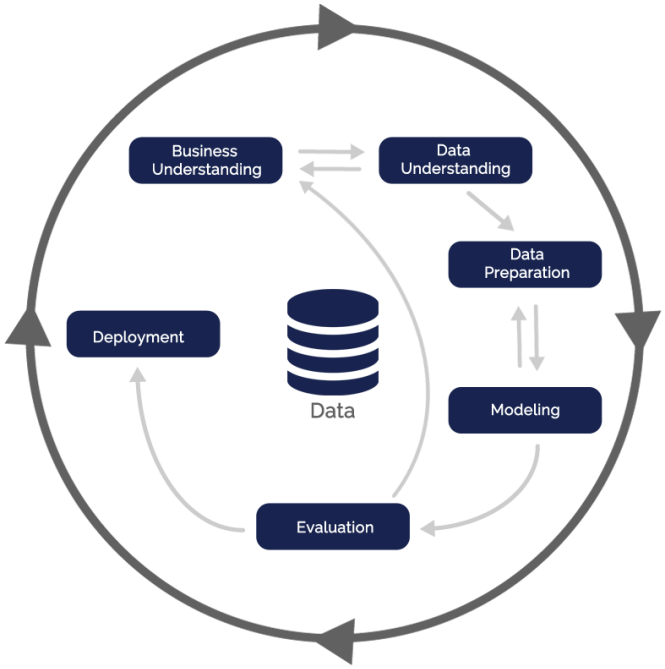


Figure 1 – CRISP-DM Diagram

According to its framework, the CRISP-DM methodology consists of six phases, which are:

1. Business Understanding: the objectives of the project are defined, and the data mining problem is formulated. This phase also involves identifying the data sources and understanding the business context of the problem.
2. Data Understanding: the data is collected, and its quality is assessed. The data is then explored to gain an understanding of its structure, distribution, and relationships.
3. Data Preparation: the data is cleaned, transformed, and formatted and the variables are selected to prepare them for modeling.
4. Modeling: various modeling techniques are applied to the data to create a predictive model.
5. Evaluation: the models are evaluated and compared by using several performance metrics.
6. Deployment: the model is integrated into the business process, and the results are communicated to stakeholders.

3.1.1. Business Understanding

This study is applied in the city of Lisbon, the capital city of Portugal, and one of the most important urban centers in Europe. With a population of over 500,000 inhabitants according to the preliminary results of the 2021 Portuguese census conducted by the National Institute of Statistics (INE), Lisbon is a bustling city that is home to a rich cultural heritage and a diverse range of economic activities. Lisbon is a popular tourist destination, being among the top tourist destinations in the world, with millions of visitors arriving each year. Regarding transportation, Lisbon boasts an extensive public transportation system that encompasses buses, trams, and a well-established metro network. Additionally, since its inauguration in 2017, Lisbon has implemented a successful bike-sharing program called GIRA, comprising around 1,410 bicycles located at 140 stations throughout the city. While Lisbon's car ownership rate is roughly 30% of households that possess a vehicle.

The issue of illegal parking is a significant concern in urban environments, including Lisbon. It not only contributes to traffic congestion and decreased road safety but also leads to revenue loss for the city and inconvenience for residents and visitors. Understanding the factors influencing illegal parking occurrences and being able to forecast the future frequency of such events can aid in effective resource allocation, enforcement strategies, and urban planning efforts.

The findings of this research, along with the Power BI report, will be valuable to multiple stakeholders involved in managing parking and traffic-related issues in Lisbon. These include:

- Lisbon Municipal Police: By understanding the underlying patterns of illegal parking behavior, the Lisbon Municipal Police can allocate their resources more effectively, plan enforcement operations, and deter future violations. The insights from this study and the Power BI report will inform their decision-making process, enabling them to develop evidence-based strategies for parking management, enforcement efforts, and policy development.
- Citizens and visitors: Improved management of parking illegality can enhance the overall quality of life for residents and visitors by reducing congestion, improving road safety, and ensuring fair access to parking spaces.

3.1.2. Data Understanding

3.1.2.1. Data collection

To develop a holistic understanding of parking illegality behavior, considering both spatial and temporal dimensions, we used three different datasets.

Table 3 summarizes the attributes from the Parking illegalities records. This dataset is the primary data source for this project that includes data on past occurrences of parking illegality, spanning from January 2017 to December 2020, provided by the Lisbon Municipal Police. This dataset has a total of 89,136 illegal occurrences including timestamps, locations, and illegality types.

Table 3 - Features of abusive parking occurrences in Lisbon

Variable	Description
Road_id	ID of the road segment where the occurrence happened
Datetime	Date and time of the occurrence
Illegality class	Type of illegality from the Portuguese traffic regulations code: crosswalk, on sidewalk, conditions access, disabled, reserved, others and unknown
Detail	Detailed description of the occurrence registered by the responsible police officer
Latitude	Latitude of the occurrence
Longitude	Longitude of the occurrence
Address	Lisbon's address of the occurrence

Furthermore, the same datasets used in the study of Jardim et al. (2022), were adopted. These datasets are displayed in Table 4 and Table 5. Table 4 presents the Road and points of interest which contains information about the road network in Lisbon, more specifically 23.096 roads, including road characteristics, such as slope, speed limits, lanes and length. Additionally, it includes points of interest such as hospitals, schools, and touristic places. This dataset provides valuable context for analyzing the relationship between road characteristics, the presence of points of interest, and the occurrence of illegal parking incidents.

Furthermore, Table 5 presents the Temporal dataset which comprises temporal data from 2017 to 2020, including day, month, year, day of week and season, as well as additional information such as weather and COVID-19 data. The weather data includes factors like humidity, precipitation, temperature, sun, and wind speed, which can help identify any correlations between weather conditions and illegal parking behavior. The COVID-19 data includes only if it was a lockdown period or not, allowing for an exploration of the impact of the lockdown pandemic on parking illegality patterns.

Table 4 - Features of road and points of interest dataset in Lisbon

Variable	Description
Road_id	Unique identifier of each road segment in Lisbon
Name	Name of the road segment
Slope	Slope of the road segment in degrees
Lanes	Number of lanes in a road segment
Oneway_road	Directionality of the road, with a value of 1 indicating a one-way road, 0 indicating a non-one-way road, and -1 indicating unknown
Bike_lane	Presence of a bike lane on the road, with a value of 1 indicating that there is a bike lane, 0 indicating no bike lane, and -1 indicating unknown
Pedestrian_lane	Presence of a sidewalk on the road, with a value of 1 indicating that there is a sidewalk, 0 indicating no sidewalk, and -1 indicating unknown
Velmax_50	Flag identifying if the road segment has maximum velocity of 50km/h or not
Vel_max	Maximum velocity allowed in a road segment
Length	Length of the road segment in meters
Lightposts	Number of light posts in the road segment
Traffic_lights	Number of traffic lights in the road segment
Bus_stations	Number of bus stations in the road segment
Metro_train	Number of metro and train stations in the road segment
Parking_spaces	Number of parking spaced in the road segment
Schools_universities	Number of schools and universities in the road segment
Tourism_places	Number of tourism places in the road segment

Table 5 - Features of temporal dataset

Variable	Description
Datetime	Date and time in the format of yyyy/mm/dd hh/mm/ss
Date	Date in the format of yyyy/mm/dd
Hour	Hour of the day
Time_slot	Time slot of the day: Dawn, Morning, Lunch, After-work
Period	Time period of the day: Dawn, Morning, Lunch, Afternoon, Night
Period2	Hour period of the day: [12 pm; 4am], [5am; 7am], [8am; 10am], [11am; 1 pm], [2 pm; 5 pm], [6 pm; 11 pm]
Day of week	Number of the day of the week
Is_Sunday	Flag identifying if it is Sunday or not
Month	Month of the date
Is_summer_spring	Flag identifying if it is summer or spring or not
Is_weekend_holiday	Flag identifying if it is weekend or not
Year	Year of the date
Is_2020	Flag identifying if it is year 2020 or not
Avg_temp	Average temperature in degrees Celsius (hourly average)
Avg_hum	Average humidity in % (hourly average)
Avg_precip	Sum of precipitation by period in mm/h (hourly average)
Avg_windspeed	Average wind speed in km/h (hourly average)
Covid_lockdown	Flag identifying if it is covid lockdown or not

In addition to the provided datasets, we collected external supplementary data encompassing various aspects such as Health, Mobility, Cultural, Leisure, Tourism, Traffic, COVID, Holidays, and Air Quality. The details and methodology for gathering this data will be further explained in chapter 3.1.3.

3.1.2.2. Exploratory data analysis

Exploratory data analysis techniques were applied to gain a deeper understanding of the data and uncover any patterns, trends, or relationships that can contribute to achieving the research objectives. At this stage, we started by doing some descriptive statistics and then we used several visualizations to represent the data in an easier and more comprehensive way such as charts, graphs, and maps.

The heat map represented in Figure 2 allow us to understand how concentrated the occurrences were within the city of Lisbon and, as a matter of example, it showed that the central areas of Lisbon and airport surrounding areas were the places with more illegalities.

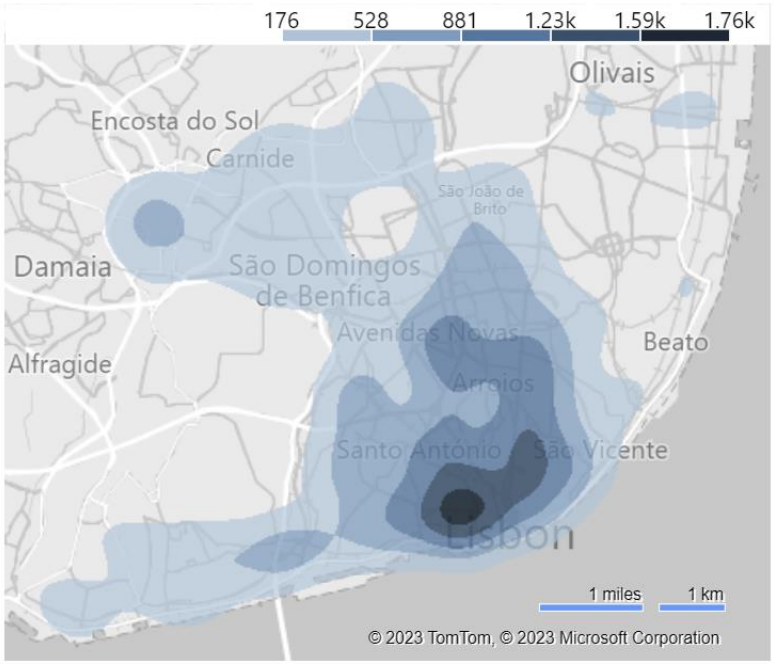
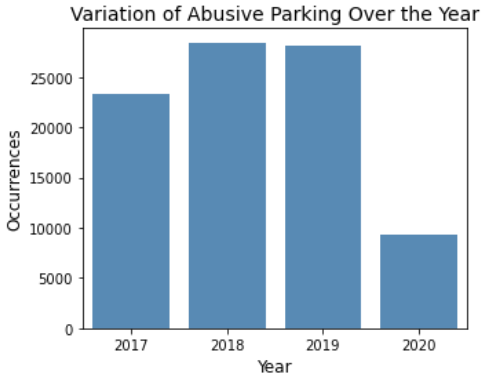


Figure 2 - Heat Map of illegal parking occurrences in Lisbon

According to Figure 3, there has been a concerning surge in the number of instances of unauthorized parking in Lisbon over the past few years, with a significant increase between 2017 and 2019 more specifically a growth of 20.59%. However, in 2020, there was a sharp decline in illegal parking cases to approximately 9000 cases (-67%), likely attributed to the COVID-19 pandemic. In addition, when we analyze the patterns on a monthly basis, February, March, and October exhibit the highest incidence of unauthorized parking, while April and August exhibit the lowest incidence.

(a) Over the years (2017 to 2020)



(b) Over the months

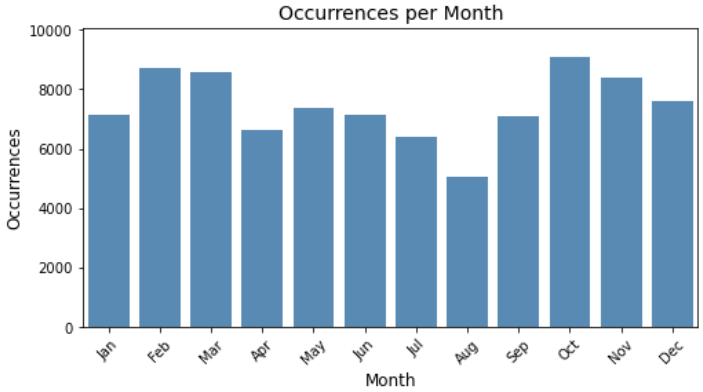


Figure 3 – Evolution of illegal parking occurrences over the years, 2017 to 2020, (a) and over the months (b)

In analyzing the occurrences of illegal parking and studying their temporal distribution, a key aspect to consider is the variation by time of day and day of the week. To gain insights into these patterns, the heat map in Figure 4 was generated, where different colors indicate the intensity of illegal parking occurrences at various time intervals throughout the day.

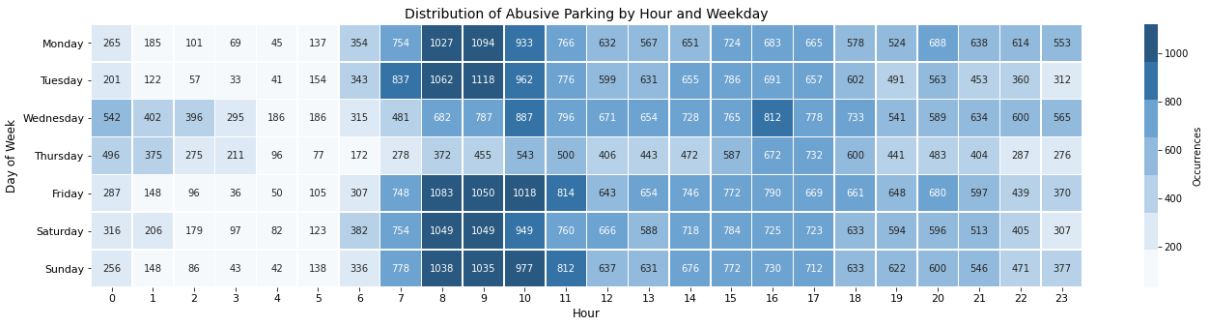


Figure 4 - Distribution of illegal parking occurrences by hour and day of the week

Illegal parking occurrences are more concentrated between 08:00 to 11:00 and less concentrated between 13:00 to 17:00. Additionally, when examining the weekdays individually, it is evident the intensity and consistency of occurrences of abusive parking, indicating a persistent issue. Among the weekdays, Saturday stands out with the highest number of recorded occurrences, and, in contrast, Sunday exhibits the lowest concentration of incidents, suggesting a relatively lower frequency of illegal parking violations on this day.

(a) over the day

(b) over the week

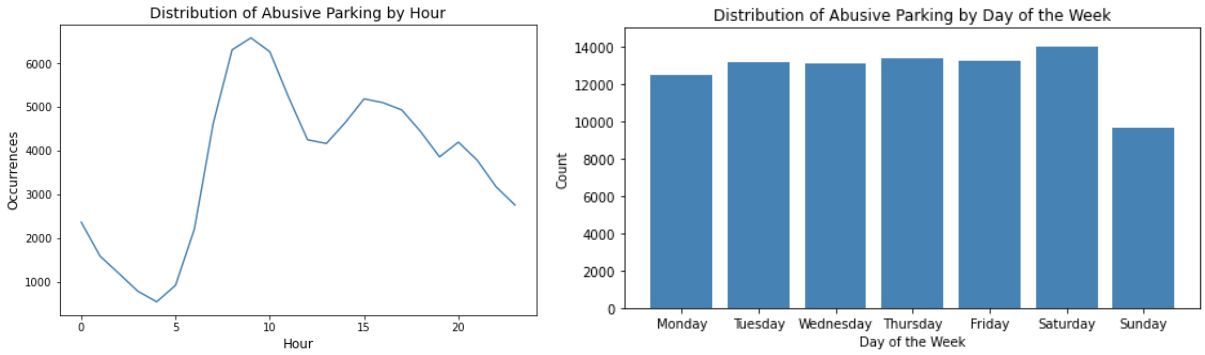


Figure 5 - Distribution of occurrences over the day (LHS) and over the week (RHS)

Taking a broader view, we can examine the variation of illegal parking incidents throughout multiple years. Figure 6 provides insights into this temporal distribution such as presenting periods like Easter and summer vacations with a lower occurrence of abusive parking. Additionally, towards the end of the graph, we observe a further decrease in incidents, which can be attributed to the impact of COVID-19 lockdown measures. This wider perspective allows us to identify the seasonal fluctuations in illegal parking behavior, highlighting the influence of vacation seasons and external factors such as pandemic-related restrictions on parking patterns.

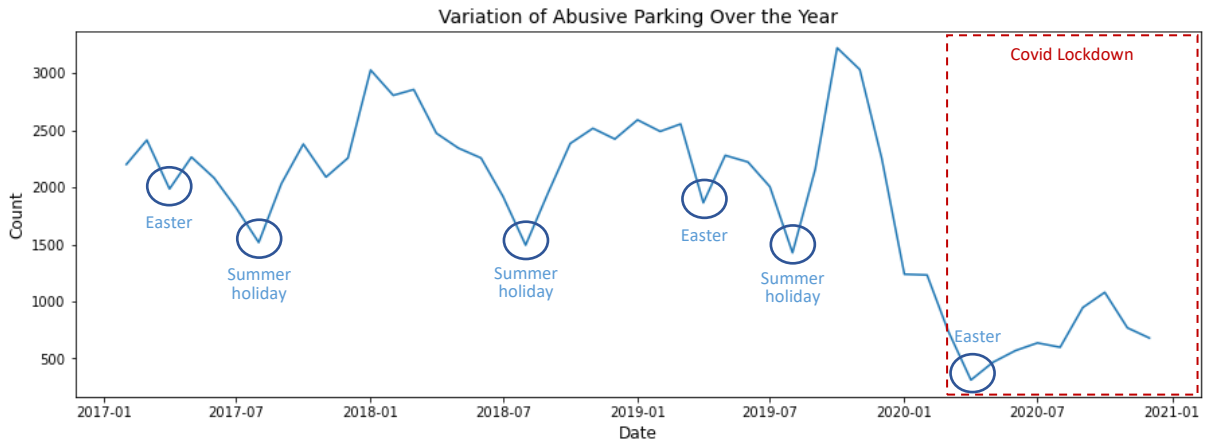


Figure 6 - Evolution of Illegal Parking Occurrences over 2017 to 2021 with Holidays and COVID-19 Lockdown Impact

Focusing on the types of illegal parking occurrences, Figure 7 shows that 40% of these incidents fall under the category of access conditioning, while 30% are attributed to violations in reserved parking places. The remaining 30% encompass a range of occurrences, including parking on sidewalks, crosswalks, spaces designated for disabled individuals, and other undefined or miscellaneous instances.

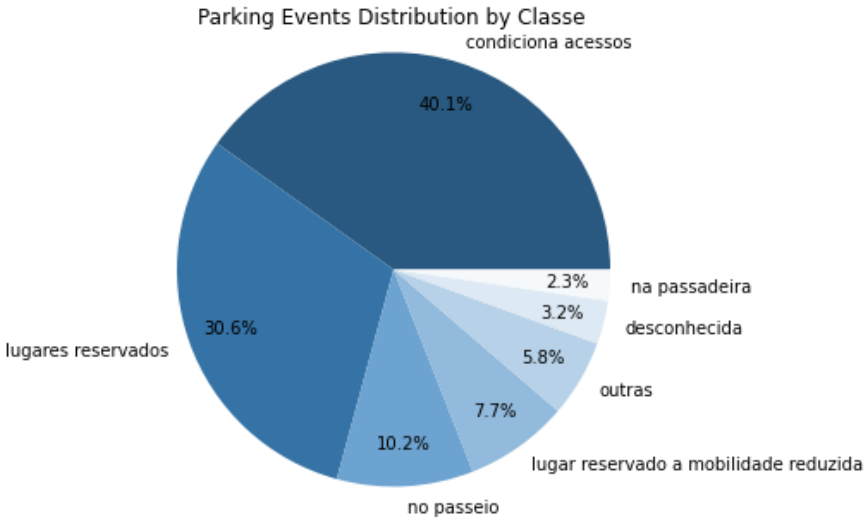


Figure 7 - Distribution of parking illegalities by Illegality Type

3.1.3. Data Preparation

The Data Preparation phase involves transforming and preparing the raw data into a suitable format for analysis. This phase ensures that the data is reliable, consistent, and properly structured to obtain meaningful insights.

3.1.3.1. Data Acquisition

In this way, in addition to the datasets mentioned in section 3.1.2., supplementary data from external sources were collected and joined to the previous datasets to enhance the analysis. In Table 6 we can see the type of data that were added, the name of the variables and their sources.

Table 6 – Features gathered and joined to the main datasets.

Datasets	Topic	Variables	Description	Source
Roads_POI	Health	Hospitals	Number of hospitals (includes private, public, and health centers)	LA
	Mobility	Bike_stations	Number of bike stations	LA
	Cultural	Cinemas_theatres	Number of cinemas and theatres	LA
		Museums	Number of museums	LA
	Leisure	Shopping_centers	Number of shopping centres	LA
	Tourism	Accommodations	Number of accomodations	LA
Parking llegalities	Traffic (Waze Alerts & Jams)	Alerts_count	Number of alerts	CML
		Jams_count	Number of jams	CML
		jam_delay	Jam's delay, in minutes (in case of block, -1)	CML
		Jam_length	Jam's length in meters	CML
		jam_speedKMH	Speed on jammed segments in Km/h	CML
		jam_level	Traffic congestion level (0 = free flow 5 = blocked)	CML
Temporal	COVID	Covid_cases	Number of new daily covid cases	WHO
	Holidays	Is_holiday	Flag identifying if it is holiday or not	CALENDARR
		Holiday_name	Name of the holiday	CALENDARR
	Air Quality	PM10	Partículas < 10 µm (µg/m3)	QualAR
		NO2	Dióxido de Azoto (µg/m3)	QualAR
		CO	Monóxido de Carbono (mg/m3)	QualAR
		SO2	Dióxido de Enxofre (µg/m3)	QualAR
		O3	Ozono (µg/m3)	QualAR
		PM25	Partículas < 2.5 µm (µg/m3)	QualAR

In order to integrate the collected data with the Roads_POI and Parking Datasets, a spatial join was employed. This process involved associating the data points from the collected data with the nearest road segment within the Lisbon Municipality. By mapping the coordinates of each data point, we were able to determine the specific road on which it is located, therefore, column “road_id” was added to each collected dataset.

For the health and cultural data, after using the previous explained mapping process, some files were concatenated into a unique dataset, i.e., “Public_hospitals”, “Private_hospitals” and “health_centers” into “hospitals” while “cinema” and “theather” were concatenated into “cinemas_theatres” due to their similarities in terms of their purpose as entertainment venues. Regarding the remaining datasets on Mobility, Leisure, and Tourism, no data concatenation was necessary as each dataset for each topic became a single variable in the Roads Dataset. Furthermore, the dataset was grouped by the "RoadID" column which allowed the creation of the new column that calculates the count of instances for each road and then we added this new column in the Roads & POI dataset through a left merged based on the "RoadID" column. Lastly, any missing values (NaN) in the new count column were replaced with 0.

As a matter of example, in the bike_stations dataset, (1) first we added the coloum Road_id through the spatial join, (2) we grouped the dataset by road_id, (3) created the new column “bike_stations” that counts the number of bike_stations per road segment and then (4) we added this new variable into Roads & POI dataset through the merge with bike_stations dataset.

For the traffic data, two types of data were provided: (1) traffic alerts incidents reported by Waze users through the Waze mobile application, and (2) Traffic Jams information that included data gathered in real time about traffic slowdowns. For each of these types, several csv files were imported and concatenated ending up with “alerts” and “jams” datasets. Additionally, for each dataset we first filtered for country equal to Portugal, then for City containing the word “Lisbon” and finally for all the traffic data between 2017 and 2020. In this way, the 2 final datasets about traffic were only regarding Lisbon city during the period time of our study. Then, we added the column “road_id” to each dataset through the performance of the spatial join, grouped the dataset by road id, date and hour and created the columns presented in Table 6, using mainly the average function. At the end, these variables were integrated into the Parking Dataset through a join operation based on the common identifiers "road_id”, “date” and “hour”. To conclude, given the fact that corresponding data from both datasets only encompass the years 2019 and 2020 and no data was available for periods outside of the specific timeframe (2017 to 2018), we found that all these variables exhibited missing values.

Regarding the Air Quality Data, we gathered from QualAR four zip files, each one for each year (from 2017 to 2020) containing the data from six different stations in Lisbon: Avenida da Liberdade, Beato, Entrecampos, Olivais Restelo and Santa Cruz de Benfica. Then, the following steps were followed:

- Import the data of each zip files through iterations over each Excel file;
- Create four datasets for each zip file, i.e., one dataset of air quality data for each year;
- Group each dataset by date calculating the average values for all columns (Partículas < 10 µm, Dióxido de Azoto, Monóxido de Carbono, Dióxido de Enxofre, Ozono and Partículas < 2.5 µm);
- Concatenate all the four datasets into a single Dataset called "air_quality_total" resulting in only one dataset regarding the air quality data for the 4 years.

Subsequently, the "air_quality_total" dataset was merged with the Temporal Dataset based on the "date" column. During this merging process, it was observed that certain variables such as “Partículas < 10 µm”, “Dióxido de Enxofre”, and “Partículas < 2.5 µm” contained missing values. These missing values occurred because certain stations did not have data available for specific dates.

Furthermore, for the Holidays data, we imported the csv file with all the 79 holidays in Lisbon from 2017 to 2020 and merged the Temporal data and "holidays" datasets based on the "datetime" column. Then we filled the null values in the newly created "is_holiday" column with zero “0”, indicating non-holiday dates and filled in the "holiday_name" column with the name of the holiday, if applicable, and left it empty for non-holiday dates.

Finally, concerning the covid data, we imported the csv file with all the data regarding covid pandemic, we filtered for country equal to Portugal and for the data until December of 2020. Finally, we merged the Temporal data and "holidays" datasets based on the "datetime" column adding the column “new_cases” to the temporal Dataset and filling the missing values of this column with “zero” indicating the not presence of covid-19 in Portugal.

3.1.3.2. Data Integration

In order to predict the number of illegal parking behavior considering the ten most critical roads of Lisbon, the period of the day and the type of day, we grouped the Parking illegalities Dataset by road_id, hour_period and is_weekend_holiday columns and created the new variable “nr_illegalities” that counts the number of illegalities for each hour in every day for each road segment in Lisbon. Additionally, we filtered this grouped dataset keeping only the rows related to the 10 streets with highest number of illegalities between 2017 to 2020 in Lisbon,

Finally, we merged the previous grouped dataset with Temporal dataset through Date and hour columns and subsequently merged with Roads & POI Dataset through the road_id column. This process resulted in a final Dataset with 122 rows and the following variables: road_id, hour_period. is_weekend_holiday, jams_count, waze_delay, waze_length, waze_speedKMH, waze_level, alerts_count, avg_temp, avg_hum, avg_precip, avg_windspeed, PM10, PM25, NO2, SO2, O3, CO, air_quality_index, name, slope, lanes, oneway_road, bike_lane, pedestrian_lane, velmax_50, vel_max, length, lightposts, traffic_lights, bus_stations, metro_train, bike_stations, parking_spaces, schools_universities, tourism_places, hospitals, cinemas_theatres, museums, accommodations, shopping_centers

3.1.3.3. Data Preprocessing

Before starting the data preprocessing itself, the dataset was divided into training and validation datasets in an 80/20 ratio to avoid overfitting in unseen data, to assess the performance of our predictive models, and consequently to have more reliable and effective predictions.

To ensure proper data consistency and avoid any potential data leakage, the data preprocessing phases were performed individually on both datasets, firstly on the training dataset and then separately to the test dataset.

3.1.3.3.1. Correlations

Multicollinearity can affect the performance and interpretation of various regression-based algorithms such as Linear, Lasso and Ridge Regressions. Therefore, to check if there is multicollinearity, i.e., high degree of correlation between predictor variables in a predictive model we looked at the correlations heatmap of the numerical variables.

In this way, we defined a threshold of 0.8 (or equal/below -0.8) to consider that two variables are highly correlated, resulting in pair of variables displayed in Table 7. However, these variables were only treated, to ensure the stability and reliability of your predictive model, later in the feature selection process.

Table 7 – Pair of Variables with Correlations of 0.8 or higher, or -0.8 or lower

Highly correlated Pair of Variables	Correlation
PM10 – PM25	0.9474
NO2 - CO	0.8277
oneway_road - bike_lane	0.9038
oneway_road - bus_stations	0.8575
bike_lane - bus_stations	0.9638
Lightposts - traffic_lights	0.9952
jams_count - waze_delay	0.9852
Jams_count - waze_level	0.9636
waze_delay - waze_length	0.9095
waze_delay - waze_speedKMH	0.8321
waze_delay - waze_level	1.0000
waze_length - waze_speedKMH	-0.8454
waze_length - waze_level	-0.8454
waze_speedKMH - waze_length	0.8502

3.1.3.3.2. Outliers

To optimize the performance of our models and account for the sensitivity of certain models to outliers (data points that deviate significantly from the overall pattern), we conducted an analysis to identify and evaluate the presence of these extreme values in our variables. Firstly, we examined the skewness and kurtosis of our variables, which provided insights into the shape and symmetry of their distributions, and then visualization tools, such as boxplots or histograms.

During our analysis, we observed a minor presence of outliers in certain variables. While there are various approaches to mitigate the influence of outliers, we made the decision not to treat them. This decision was based on two factors: the dataset's small size and our thorough examination confirming that the outliers were not inconsistencies or errors. Instead, we focused on applying appropriate transformations, as explainer further in this section, to address skewed variables and achieve more symmetrical distributions, thereby minimizing the impact of these extreme values.

3.1.3.3.3. Missing Values

Having the presence of missing values in our database is a problem that, if not detected correctly, can lead to biased or unreliable results and, consequently, leading to wrong assumptions and decisions.

Before adding new variables, to do the spatial join and adding the column “road_id” in the imported datasets, firstly we needed to treat the missing values from the Parking illegalities Dataset. This dataset presented missing values in the following variables:

- Road_id, Latitude, Longitude (107 values)
- Name (2461 values)

Regarding the first three variables, since their rows without values coincide and the proportion is not significant (0.12% of data), we opted to delete them. Then, for the variable name, the missing values can be attributed to errors in the road dataset, where certain road segments had incorrect names that resulted in missing values upon joining the datasets. To guarantee consistency, we filled these missing values with "unknown" in the parking dataset and also replaced the erroneous values of "#NAME?" with "unknown."

Additionally, after incorporating new variables into our datasets, the variables related to traffic information presented a significant proportion of missing values from the dataset of approximately 25% as we can see in Figure 8. This was already expected since, as mentioned earlier in the 3.1.3.1 section, the traffic data was only available for the years 2019 and 2020. In this way, we opted to delete these four variables.

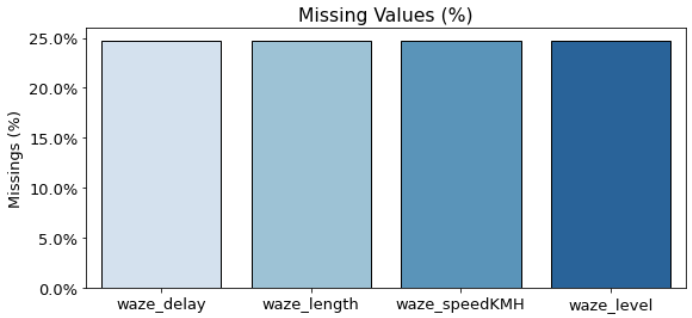


Figure 8 - Variables with missing values (% of total)

3.1.3.3.4. Feature engineering

To potentially increase the accuracy, efficiency, and clarity of our data, we transformed existing variables and created new ones based on the originals.

Table 8 – Variables' transformation and creation

Transformations	Variable	Description
Simple	Air_quality_index	Categorical variable based on the Lisbon Environmental Index Emission Criteria and on the previous air quality variables. Categories: "Bom", "Médio" and "Fraco"
Logarithmic/sqr/sqr3	avg_precip_log	Logarithm of the variable avg_precip
	avg_precip_sqrt	Square Root of the variable avg_precip
	avg_precip_sqrt3	Cubic Root of the variable avg_precip
	tourism_places_log	Logarithm of the variable tourism_places
	tourism_places_sqrt	Square Root of the variable tourism_places
	tourism_places_sqrt3	Cubic Root of the variable tourism_places
	slope_log	Logarithm of the variable slope
	slope_sqrt	Square Root of the variable slope
	slope_sqrt3	Cubic Root of the variable slope
	CO_log	Logarithm of the variable CO
	CO_sqrt	Square Root of the variable CO
	CO_sqrt3	Cubic Root of the variable CO
	NO2_log	Logarithm of the variable NO2
	NO2_sqrt	Square Root of the variable NO2
	NO2_sqrt3	Cubic Root of the variable NO2
Encoding	Hour_period	Hour period of the day: [12 pm; 3am] = 1, [4am; 6am] = 2, [7am; 9am] = 3, [10am; 1pm] = 4, [2 pm; 4pm] = 5, [5pm; 8 pm] = 6, [5pm; 8pm] = 7
	air_quality_index_Code	Air quality index code: 'Fraco' = 1, 'Médio' = 2, 'Bom' = 3

Firstly, the Air_quality_index variable was created based on the Lisbon Environmental Index Emission Criteria developed by the Lisbon City Council (CML). This index criteria assigns a color to the hazard/intensity level of the air quality parameters graded in green for normal situations, yellow, orange and red. In this way, we created an index for each air quality parameter with the categories good for the green ones, medium for yellow and bad for red. Then, the overall "Air_quality_index" variable was created which represents the worst index of all the variables, i.e., if all the variables are "Good" but one of them is "Bad" the overall index at that time will be "Bad".

In order to mitigate the influence of skewed variables and achieve more symmetrical distributions, we employed transformations such as logarithm, square root, and cubic root on variables that exhibited

significant skewness, namely avg_precip, tourism_places, slope, CO, and NO2. While the symmetry of variable distributions is not a strict requirement for all predictive algorithms, algorithms such as Linear Regression, KNN, neural networks that are implemented in this study can benefit from symmetric distributions or are more robust to non-symmetric data.

Finally, to include the categorical variables like Hour_period and air_quality_index in our model, we converted them into numerical by resorting through encoding, more specifically to the Ordinal encoding where it keeps the natural order of the variables.

3.1.3.3.5. Scaling Data

At this stage, our dataset contains numerical data that varies across different scales, which can lead to misleading interpretations in most of the predictive algorithms developed in this project, as variables with larger scales may have a false larger impact compared to those with smaller scales. To address this issue and ensure accurate comparisons among the variables, we normalized the impact of each numerical variable by transforming them into the same scale using MinMax Scaler, i.e., into a range between 0 and 1.

3.1.3.4. Feature selection

Before we started the modeling part, we performed feature selection by selecting a smaller subset of variables that could better explain the target variable and reduce overfitting while improving model accuracy. To make this selection, we utilized the following distinct techniques: (1) Correlations (Pearson and Spearman), (2) Decision Trees, (3) Random Forest, (4) Gradient Boosting (5) RFE, (6) Lasso and (7) Step-Forward selection to make better decisions on which variables to keep.

Then, after performing all the approaches, we combined their results in a table to decide which variables to eliminate from our model through a defined threshold, i.e., eliminate the variables that were chosen to be excluded in at least 4 approaches. Additionally, we checked the correlation of the variables that were excluded in at least three approaches with the target and kept the ones with high coefficients.

3.1.4. Modeling

At this phase, the predictive algorithms shown in Table 9 were developed and tested with parameter tuning techniques to predict illegal parking occurrences. It includes Ensembles methods since they provide a robust solution by leveraging the collective knowledge of multiple models instead of relying on a single model, improving generalization, and enhancing predictive performance.

Table 9 – Description of the predictive models developed

Models	Description	Reference
Linear Regression	Simple and interpretable baseline model since it understands the direct impact of each independent variable on the target variable. In addition, it assumes linearity and independence of errors in the model.	(Weisberg, 2005)
Decision Tress Regression	The simplest tree-based model that can capture non-linear relationships and interactions among variables, providing flexibility in modeling complex patterns. Enables the identification of important features, handles missing data, and requires less data preprocessing. However, it cannot deal with correlated variables, and it is highly prone to overfitting.	(Apté & Weiss, 1997)
Lasso Regression	Also known as L1 regularization, is a linear regression technique that performs both variable selection and regularization and uses a penalty term to encourage sparsity in the coefficient estimates, effectively shrinking some coefficients to zero. It controls model complexity and prevents overfitting, resulting in a more robust model.	(J. Friedman et al., 2010)
Ridge Regression	Also known as L2 regularization, is a linear regression technique that unlike Lasso Regression, does not perform feature selection but instead, reduces the impact of correlated variables effectively handling multicollinearity issues and reducing their impact on model performance.	
ElasticNet Regression	Combines L1 and L2 regularization techniques, offering a flexible approach to handle multicollinearity while encouraging sparsity. Usually, it is effective in situations where there are both strong and weak predictors.	
Bayesian Regression	Enables the incorporation of prior knowledge, provides a distribution of parameter estimates, and facilitates uncertainty analysis. This approach is particularly useful when dealing with limited data or when prior knowledge is available for informative prior specification	(Minka, 2000)
Neural Networks	Powerful in capturing complex patterns and interactions within the data, particularly useful when non-linear relationships are present. Training deep networks may require more data and computational resources.	(Wang, 2003)
KNN	It is s a non-parametric method that utilizes the similarity between observations to make predictions, allowing the model to incorporate the characteristics of similar instances. Requires appropriate scaling and consideration of the optimal number of neighbors.	(Peterson, 2009)
Ada Boosting	Ensemble technique that combines multiple weak learners to create a strong predictive model. It assigns weights to each training sample, focusing more on misclassified samples in subsequent iterations.	(Dietterich, 2000)
Bagging	Ensemble method that creates subsets of the data and fits multiple models independently, reducing variance and increases stability by incorporating diverse models.	
Random Forest	Ensemble method that builds multiple decision trees using bootstrapped samples and random feature subsets, resulting in improved generalization and robustness.	(Breiman, 2001)
Gradient Boosting	Ensemble method that iteratively improves weak models by focusing on the residual errors, leading to improved accuracy and the ability to capture complex relationships in the data	(J. H. Friedman, 2001)

3.1.5. Evaluation

Following the modeling part, we evaluated and compared the performance of the predictive models using appropriate evaluation metrics such as Mean Squared Error, Root Mean Squared Error, and R-squared to select the best-performing model for our study.

Table 10 - Regression Model Evaluation Metrics and Interpretations

Metric	Interpretation
MSE	Lower MSE indicates better fit of the model to the data, i.e., smaller errors between the predicted and observed values
RMSE	Lower RMSE indicates better fit of the model to the data, similar to MSE
R2	Higher R-squared values indicate a better fit of the regression model to the data, ranging between 0 and 1

3.1.6. Deployment

In this phase, the model is integrated into the business process, and the results are communicated to stakeholders. In order to communicate the insights in an understandable and visually appealing manner, we connected our data to PowerBI tool and developed dashboards according to each business area. We opted to create and present the following dashboards:

- Overview dashboard, the initial dashboard that presents a high-level summary of key metrics and trends related with illegal parking data in a quick and intuitive way.
- The evolution of illegalities dashboard, which is aimed at uncovering patterns related to illegal parking occurrences over time, providing insights into which months and days of the week are more susceptible to illegal parking incidents. This dashboard also enables an analysis of the evolution of illegal parking incidents over the years and can help identify any potential spatial metrics that may be contributing to this behavior.
- Classes of illegalities dashboard, which provides a comparative examination of the various types of unauthorized parking, as well as an analysis of each type of illegality individually.

The dashboard creation process started by uploading the datasets already preprocessed in Python into the Power Bi. Then, the tables' structure and relationships were created ending up with the multi-dimensional data model displayed in Figure 9.

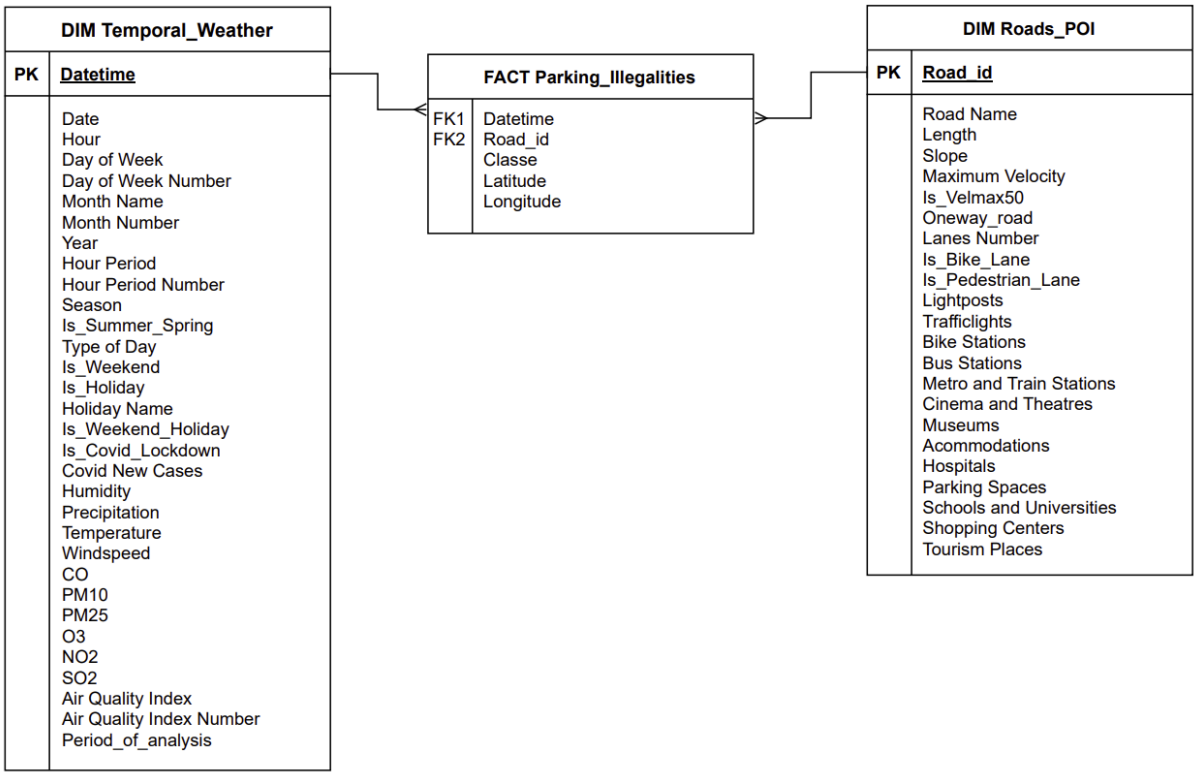


Figure 9 – Multidimensional Data Model

4. RESULTS AND DISCUSSION

4.1. FEATURE SELECTION

4.1.1. Correlations

The first approach we used for feature selection was the analysis of the correlations between all the variables, i.e., to check the existence of multicollinearity in our model. For that, we looked at the correlation matrixes using the Spearman and Pearson Coefficient and we concluded that there were several highly correlated variables with coefficients of 0.8 or higher (or -0.8 or lower), which were expected since those pairs of variables have one variable that was created based on the other, such as the logarithmic and Boolean ones.

To address the issue of correlated variables, we adopted a two-step approach. Firstly, we assessed the correlation of each variable with the target variable (nr_illegalities) and retained the one with the higher coefficient. However, in cases where the correlated variables had identical coefficients with the target, we employed Decision Tree bar plots to examine feature importance and prioritized variables with higher importance.

For the following feature selection approaches, we did not use the correlated variables as they can affect the performance of some of those methods.

4.1.2. Decision Trees and Random Forest

Then, we employed decision trees to evaluate the importance of each feature based on their position within the tree structure, i.e., considering the proximity of features to the root of the tree, and Random Forest, which inherently provides rankings of feature importance, enabling us to identify the most influential features in the model's predictions.

To accomplish this, we generated bar plots for Decision Tree and Random Forest models displaying the variables in order of importance and decided, based on a reasonable number of variables desired, to select as candidates for elimination the least important ones. Regarding the Decision Trees, we were able to identify the following variables to be included: road_id, avg_temp, avg_hum, avg_windspeed, SO2, slope, bus_stations, avg_precip_sqrt3, NO2_log, hour_period, parking_spaces and PM10. Similarly, for the Random Forest, the following variables were included: road_id, is_weekend_holiday, tourism_places_sqrt3, length, lightposts, avg_temp, avg_hum, avg_windspeed, SO2, slope, avg_precip_sqrt3, NO2_log, hour_period, parking_spaces and PM10.

4.1.3. Gradient Boosting

Additionally, the gradient boosting algorithm provides a measure of feature importance by evaluating how much each feature contributes to improving the model's performance. By examining the bar plot representing the feature importance scores, we identified the following variables with higher importance values to keep in our model: road_id, is_weekend_holiday, tourism_places_sqrt3, lightposts, avg_temp, avg_hum, avg_windspeed, SO2, slope, avg_precip_sqrt3, NO2_log, hour_period, parking_spaces and PM10.

4.1.4. Recursive Feature Elimination

To perform RFE on our features, we used the Linear Regression model, and decided to keep 22 features that gave us better Mean Squared Error. This number was determined by analyzing the elbow plot in Figure 10, which indicates the point at which the decrease in MSE significantly diminishes. This model ranked features according to their importance by iteratively eliminating the least important features and refitting the model until the desired number of features is reached.

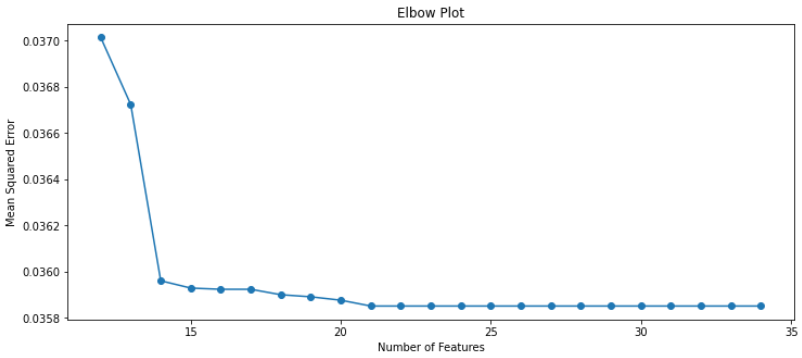


Figure 10 - Elbow Plot for RFE feature selection

Then, based on the model's analysis, the recommended variables to be retained are: road_id, is_weekend_holiday, avg_temp, avg_hum, avg_windspeed, slope, lanes, length, bus_stations, parking_spaces, accommodations, shopping_centers, avg_precip_sqrt3 and NO2_log

4.1.5. Lasso

Then, we applied the Lasso model, a regularization technique that helps in feature selection by reduction the coefficients of less important variables to zero and effectively selecting only the most relevant features. Through this approach, we were able to identify the following variables to be included: road_id, is_weekend_holiday, avg_temp, avg_windspeed, SO2, slope, lanes, length, bus_stations, avg_precip_sqrt3, NO2_log and air_quality_index_Code.

4.1.6. Step-Forward selection

Finally, we also decided to use Step-Forward Selection as a feature selection method, using the Linear Regression model. This technique starts with an empty set of features and iteratively adds the most significant variable at each step. In this way, we identified the following variables to be included: road_id, is_weekend_holiday, avg_hum, avg_windspeed, tourism_places_sqrt and NO2_log.

4.1.7. Evaluation of Feature Selection Methods

After conducting all the feature selection approaches, we consolidated their individual results into the following table to determine the variables to be excluded from our model.

To accomplish this, our initial step involved removing variables that were chosen to be excluded in at least four of the techniques and then we looked at the remaining variables that were only excluded in three and checked their correlations with the target variable. At the end, the ones with the lowest correlation coefficients were eliminated.

Table 11 - Feature Selection Table with Decision Process for all the Models

Variables / Models	DT	RF	GB	RFE	Lasso L1	Step-Forward	DECISION	Correl. w/ target	FINAL DECISION
road_id									YES
avg_precip_sqrt3									YES
SO2									YES
is_weekend_holiday									YES
hour_period							MAYBE	YES	YES
air_quality_index_Code									NO
accommodations									NO
avg_windspeed									YES
tourism_places_sqrt							MAYBE	YES	YES
length							MAYBE	NO	NO
slope									YES
shopping_centers									NO
avg_temp									YES
lightposts									NO
avg_hum									YES
parking_spaces									YES
lanes									NO
NO2_log									YES
PM10							MAYBE	YES	YES
bus_stations							MAYBE	YES	YES
jams_count									NO

To conclude, we kept 14 variables in our final dataset, i.e., the ones in Green in the Final Decision column in Table 11.

4.2. PREDICTIVE MODELS

With the objective of achieving the most accurate prediction of the *nr_illegalities* target variable, we implemented nine algorithms – Linear regression, Decision Trees, Lasso, Ridge, ElasticNet, Neural Networks, KNN, and Bayesian Regression – and four ensemble classifiers – Random Forest, Gradient Boosting, Bagging, and Ada Boosting.

Subsequently, we evaluated the performance of each model by analyzing the mean squared error scores, root mean squared error and r-square, we used techniques like GridSearchCV to identify the optimal values of hyperparameters, and at the end, we identified the best-performing model through a comparison analysis.

4.2.1. Linear Regression

Regarding the Linear Regression, we started by creating a model with the default parameters and checking the score in the train and test datasets. To achieve the best possible results, we conducted parameter tuning that resulted in the following scores available in Table 12.

- • Fit Intercept (whether to calculate the intercept): False

Table 12 – Scores of the performance metrics for Linear regression

MSE		RMSE		R2	
Training	Test	Training	Test	Training	Test
0.03721	0.06188	0.19291	0.24876	0.218	0.022

4.2.2. Decision Trees Regression

Initially, the decision tree model was fitted without specifying any parameters, leading to a model that exhibited overfitting. Consequently, hyperparameter tuning was performed in an attempt to enhance its performance.

Hence, we conducted a comprehensive analysis of each parameter individually, as depicted in Figure 11, to determine the optimal range of values that minimize overfitting and yield lower scores.

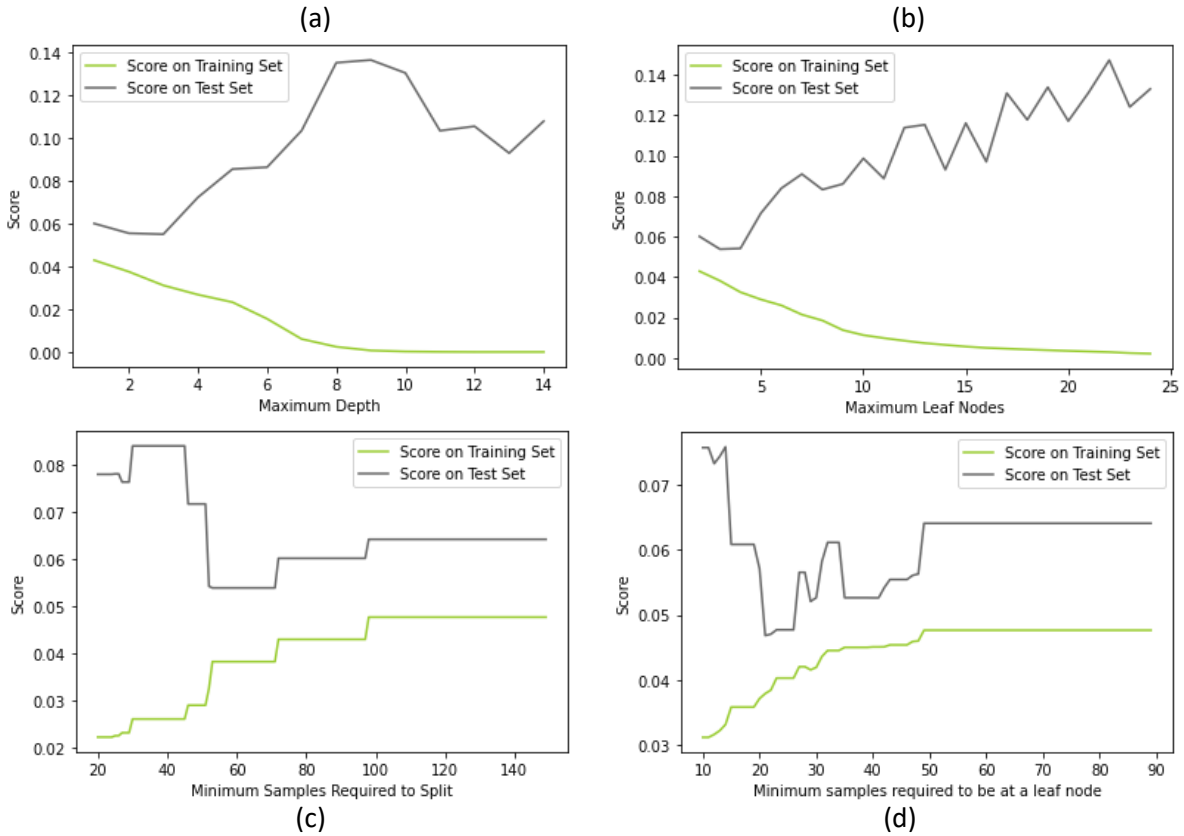


Figure 11 - Maximum depth, (a) maximum leaf nodes, (b) minimum samples required to split, (c) and minimum samples required to be at a leaf node (d) with training vs. test MSE scores

Then, we combined the 4 parameters using GridSearchCV, and obtained the following set of parameters that yielded the best performance:

- Maximum Depth: 2,
- Maximum Leaf Nodes: 3,
- Minimum Samples per Leaf: 39,
- Minimum Samples per Split: 55

When fitting the final model, with the parameters indicated above, the result was as follows:

Table 13 - Scores of the performance metrics for Decision Trees Regression

MSE		RMSE		R2	
Training	Test	Training	Test	Training	Test
0.04498	0.05264	0.21209	0.22943	0.055	0.168

4.2.3. Lasso regression

Regarding the Lasso Regression, the parameter tuning was performed using Grid Search Cross Validation, resulting in the following parameters and scores:

- Alpha (regularization parameter): 0.001,
- Maximum Iterations: 1000,
- Tolerance: 0.001.

Table 14 - Scores of the performance metrics for Lasso regression

MSE		RMSE		R2	
Training	Test	Training	Test	Training	Test
0.03750	0.05671	0.19367	0.23815	0.212	0.103

4.2.4. Ridge Regression

Using the same method, for the Ridge Regression, the parameter tuning was performed using GridSearchCV, resulting in the following parameters and scores:

- Alpha (regularization parameter): 5,
- Solver (algorithm used for optimization): saga

Table 15 - Scores of the performance metrics for Ridge regression

MSE		RMSE		R2	
Training	Test	Training	Test	Training	Test
0.03995	0.05195	0.19988	0.22793	0.161	0.179

4.2.5. ElasticNet Regression

Additionally, for the ElasticNet Regression, the parameter tuning was also performed using GridSearchCV, resulting in the following parameters and scores:

- Alpha (regularization parameter): 0.001,
- Fit Intercept: False,
- L1 Ratio: 0.7,
- Maximum Iterations: 200,
- Tolerance: 0.0001

Table 16 - Scores of the performance metrics for ElasticNet regression

MSE		RMSE		R2	
Training	Test	Training	Test	Training	Test
0.03736	0.05817	0.19329	0.24118	0.216	0.081

4.2.6. K-Nearest Neighbors (KNN)

Afterwards, the K-Nearest Neighbors Regression together with the GridSearchCV presented the following parameters and scores:

- Algorithm: auto
- Distance Metric: minkowski,
- Number of Neighbors: 9
- Power parameter for Minkowski metric (1 for Manhattan, 2 for Euclidean distance): 1,
- Weighting scheme: uniform

Table 17 - Scores of the performance metrics for K-Nearest Neighbors

MSE		RMSE		R2	
Training	Test	Training	Test	Training	Test
0.03726	0.04930	0.19303	0.22204	0.218	0.221

4.2.7. Neural Network

The Neural Network Regression jointly with the GridSearchCV presented the following parameters and scores:

- Activation function: relu,
- Alpha (regularization parameter): 0.0001,
- Hidden layer sizes: (30, 20, 10),
- Learning Rate: adaptive,
- Maximum iterations: 500,
- Solver (algorithm used for optimization): sgd

Table 18 - Scores of the performance metrics for Neural Networks

MSE		RMSE		R2	
Training	Test	Training	Test	Training	Test
0.04556	0.06960	0.21347	0.26381	0.043	0.090

4.2.8. Bayesian Linear Regression

Afterwards, for the Bayesian Regression, we opted to use the LinearRegression and perform the parameter tuning using GridSearchCV, resulting in the following parameters and scores:

- Alpha 1: 0.0001,
- Alpha 2: 1e-06,
- Lambda 1': 1e-06,
- Lambda 2: 0.0001

Table 19 - Scores of the performance metrics for Bayesian Linear Regression

MSE		RMSE		R2	
Training	Test	Training	Test	Training	Test
0.04188	0.05392	0.20466	0.23221	0.121	0.148

4.2.9. Ensemble Methods:

4.2.9.1. AdaBoost

The Adaboost model used the Decision Trees regressor stated previously, and together with the GridSearchCV presented the following parameters and scores:

- Learning Rate: 0.5,
- Number of Estimators: 10,
- Random State: 3

Table 20 - Scores of the performance metrics for AdaBoost ensemble

MSE		RMSE		R2	
Training	Test	Training	Test	Training	Test
0.04144	0.04200	0.20356	0.20495	0.130	0.336

4.2.9.2. Bagging

The Bagging model also used the Decision Trees regressor stated previously, and together with the GridSearchCV presented the following parameters and scores:

- Maximum Features: 11,
- Maximum Samples: 9,
- Number of Estimators: 4

Table 21 - Scores of the performance metrics for Bagging ensemble

MSE		RMSE		R2	
Training	Test	Training	Test	Training	Test
0.04793	0.06340	0.21895	0.25181	0.006	0.0

4.2.9.3. Random forest

The Random Forest Regression together with the GridSearchCV presented the following parameters and scores:

- Maximum Depth: 16,
- Maximum Leaf Nodes: 10,
- Number of Estimators: 11
- Minimum Samples per Leaf: 10
- Minimum Samples per Split: 2

Table 22 - Scores of the performance metrics for Random Forest ensemble

MSE		RMSE		R2	
Training	Test	Training	Test	Training	Test
0.03240	0.05332	0.17999	0.23091	0.320	0.157

4.2.9.4. Gradient Boosting

The Gradient Boosting Regression together with the GridSearchCV presented the following parameters and scores:

- Maximum Depth =3,
- Learning Rate =0.01,
- Number of Estimators =200,
- Minimum Samples per Leaf =10
- Minimum Samples per Split =10

Table 23 - Scores of the performance metrics for Gradient Boosting ensemble

MSE		RMSE		R2	
Training	Test	Training	Test	Training	Test
0.02407	0.05921	0.15516	0.24333	0.495	0.064

4.2.10. Discussion and Selection of the Optimal Predictive Model

Following the development of various predictive models including parameter optimization, a comparison of metric scores and overfitting levels was conducted to identify the model with the best performance.

In the next table, it is possible to observe the values of the performance metrics Mean Squared Error, Root Mean Squared Error, and R-squared for both the training and test datasets of all the models.

Table 24 – Predictive Model's Comparison

Method	**MSE**		**RMSE**		**R2**	
	Training	Test	Training	Test	Training	Test
Linear Regression	0.037	0.062	0.193	0.249	0.218	0.022
Decision Trees Regression	0.045	0.053	0.212	0.229	0.055	0.168
Lasso	0.038	0.057	0.194	0.238	0.212	0.103
Ridge	0.040	0.052	0.200	0.228	0.161	0.179
Elastic Net	0.037	0.058	0.193	0.241	0.216	0.081
KNN	0.037	0.049	0.193	0.222	0.218	0.221
Neural Network	0.046	0.070	0.214	0.264	0.043	0.090
Bayesian Linear Regression	0.042	0.054	0.205	0.232	0.121	0.148
Ada Boosting	0.041	0.042	0.204	0.205	0.130	0.336
Bagging	0.048	0.063	0.219	0.252	0.006	0.0
Random Forest	0.032	0.053	0.180	0.231	0.320	0.157
Gradient Boosting	0.024	0.059	0.155	0.243	0.495	0.064

Additionally, in order to enhance the comparison, Figure 12 has been included as graphical representations of the information presented in Table 24. This figure provides a visual examination of each performance metric individually, allowing for a comprehensive comparison between models and an assessment of overfitting for each model.

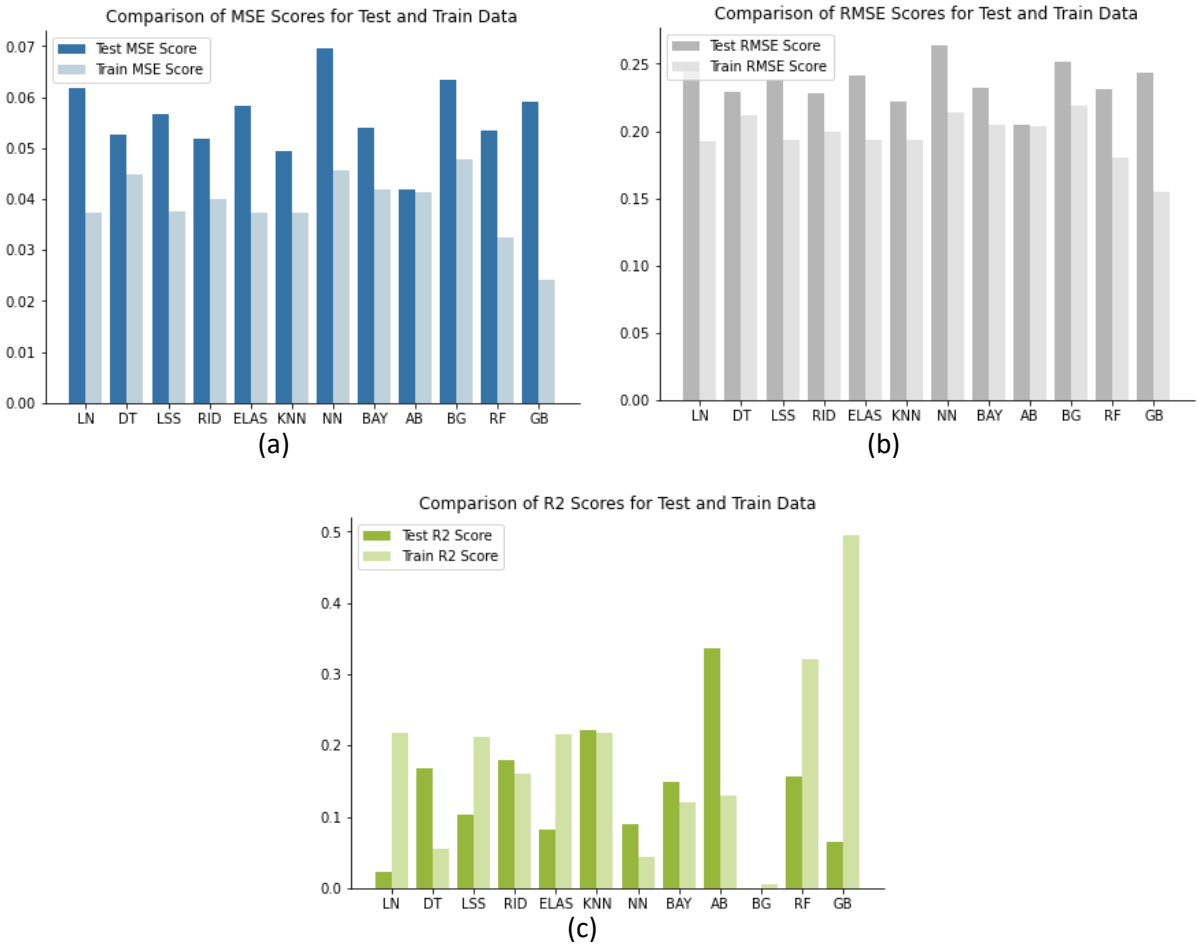


Figure 12 – Comparison of MSE (a), RMSE (b), and R2 (c) metric scores for Test and Train datasets, between all the models

Finally, to improve the comparative analysis, Figure 13 provides a visual understanding of the level of overfitting of each model in terms of the different performance metrics.

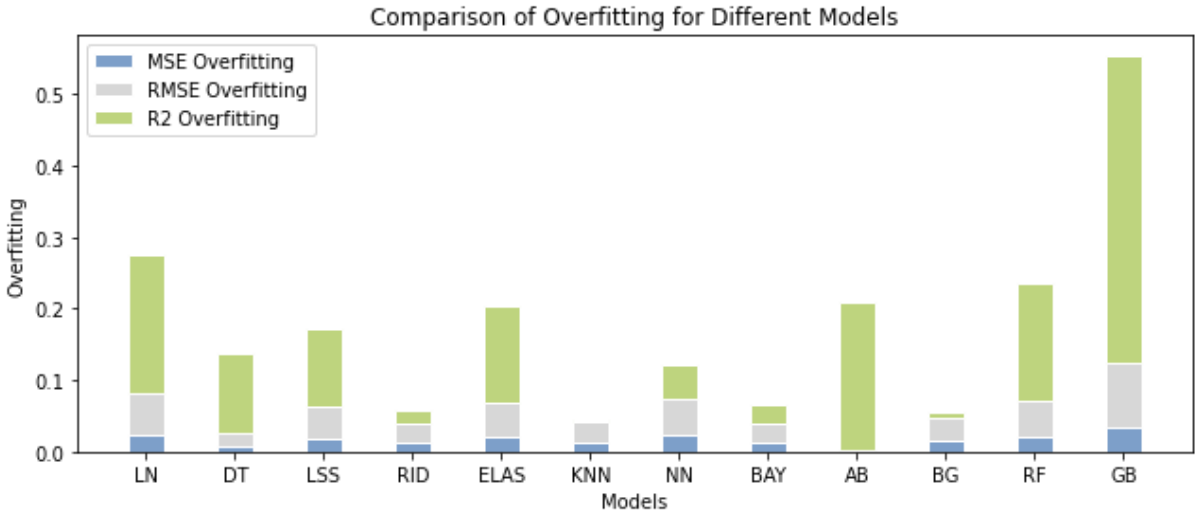


Figure 13 –Comparison of Overfitting for all the Models

Based on the previous results, KNN and AdaBoost stood out as the models that yielded the most promising results. AdaBoost demonstrated superior performance by exhibiting the lowest MSE and RMSE while effectively avoiding overfitting and a very good R-squared value in the test dataset, i.e., the model explained 36% of the variance in the target variable. However, it displayed a higher and significant level of overfitting in terms of the R-squared performance metric. On the other hand, KNN delivered outstanding results in terms of the R-squared metric without overfitting, although it exhibited some overfitting in the error metrics, albeit at low levels. These two models are very similar but since the priority of this study is precision and minimizing prediction errors, AdaBoost was selected as the optimal predictive model.

In contrast, it is also evident that Neural Networks and Bagging models exhibited the worst overall performance. These models demonstrated higher Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) values, and lower R-squared (R2) scores for both the training and validation datasets indicating poorer predictive accuracy and limited generalization capabilities.

Regarding the remaining ensemble methods, the Random Forest and Gradient Boosting, they were by far the models that achieved the best results in the training dataset, however they fail to generalize well on test data presenting the higher levels of overfitting.

Moreover, Linear Regression, Lasso, and Elastic Net models exhibited similar performance to the previous two models, albeit with smoother results. Conversely, Decision Trees demonstrated good results on the test dataset. However, it showed signs of overfitting, which was expected since Decision trees are models with high propensity to overfitting, resulting in less favorable outcomes.

To conclude, both Ridge Regression and Bayesian Regression demonstrated comparable results, characterized by low levels of errors and overfitting but not surpassing AdaBoost in terms of all the predictive accuracy.

4.3. POWER BI DASHBOARDS

As previously mentioned, one of the goals of this project is to represent illegal parking behavior in Lisbon through the creation of a Power BI dashboard. Thus, after using various pre-processing techniques we created the dashboard with the following three pages.

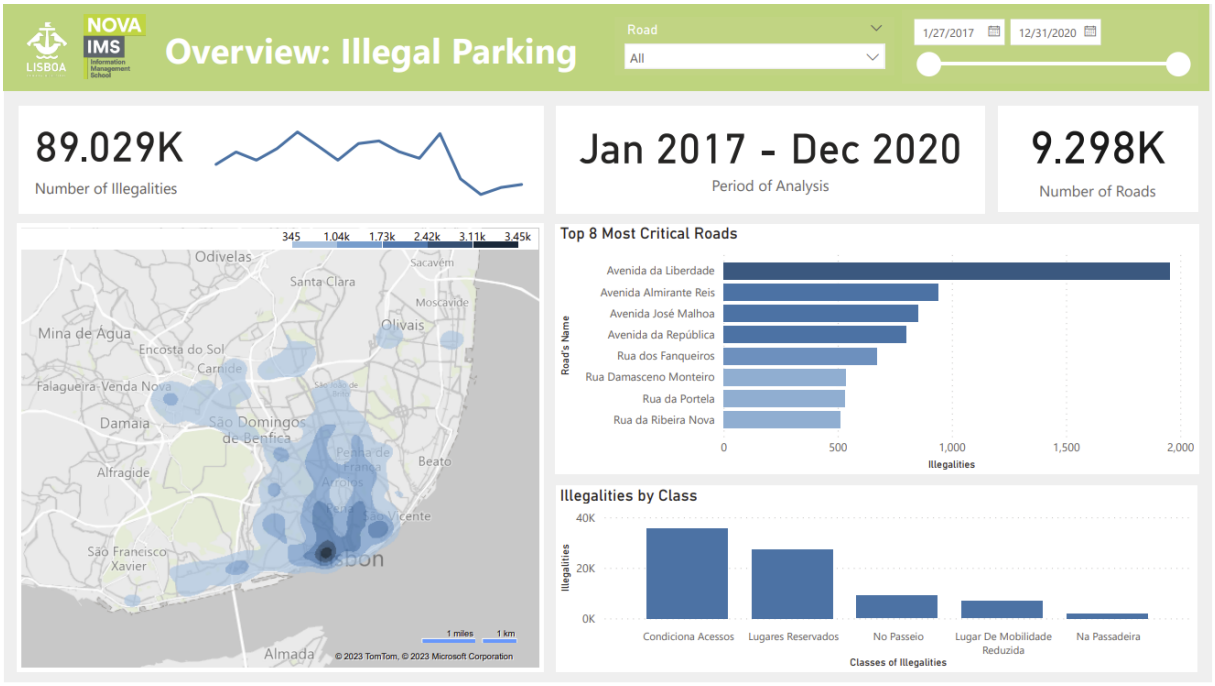


Figure 14 – Dashboard Overview of Illegal Parking

On the first page of the report, depicted in Figure 14, the Overview dashboard is showcased, providing an overview of the illegal parking behavior in Lisbon from 2017 to 2020, i.e., a high-level, concise summary of key insights related to these occurrences.

Two slicers and tree cards are available on the top of the dashboard allowing users to explore the data based on different road segments and time periods and showing the number of illegal parking occurrences and its evolution over time, the period of the analysis, and the number of road segments that occurred illegal parking occurrences. Furthermore, the heat map represents the distribution of illegal parking occurrences across Lisbon, highlighting hotspots with high illegalities rates, and the horizontal bar chart at the right show the top eight roads that are more critical, and the police should pay more attention. Finally, a vertical bar chart categorizes illegalities based on type, shedding light on the predominant types and their frequencies.

By analyzing this dashboard, it becomes evident that the central area of Lisbon, particularly near the sea, experiences the highest concentration of illegal parking occurrences, being "Avenida da Liberdade" and "Avenida Almirante Reis" the most problematic roads in this regard. Furthermore, the dashboard reveals a significant decrease in illegal parking incidents over the last year of analysis, and, indicated that the most prevalent occurrences are related to the "Condiciona acessos" and "Reservados" classes, while incidents categorized as "Na passadeira" exhibit the lowest frequency.

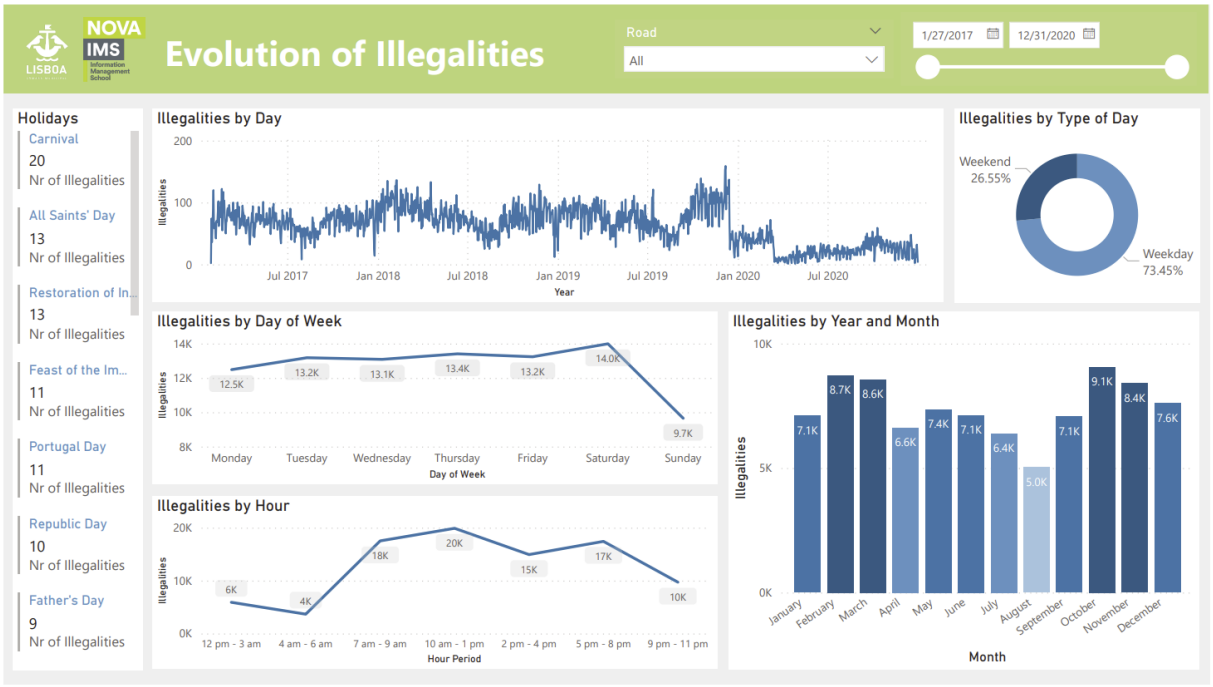


Figure 15 – Dashboard of the Parking Illegalities' Evolution

The second page represented in Figure 15, provides a comprehensive analysis of illegal parking incidents over time including the same slicers as the previous page, to explore the data based on road segments and specific timeframes. This page provides insights into temporal patterns including line graphs showing occurrences by day, day hour period, and day of the week. Additionally, it also features a pie chart categorizing illegalities by the type of day (weekend or weekday) and a bar chart displaying occurrences by month and year. Finally, a multi-row card displays the holidays in Portugal, along with the number of illegalities recorded on those specific dates, and is arranged in a decreasing order, starting with the one with the highest frequency of violations.

Regarding the temporal dimension, the most critical period of the day can be found in the period between 10 am and 1 pm, following by the period immediately before, between 7 am and 9 am, and the period between 5pm and 8pm, with values decreasing until the end of the day, being the period between 4 am and 6 am the one with the lowest number of occurrences. As for the day type, it shows a significant higher percentage of occurrences on working days than on weekends, being Saturday the most critical day of week and Sunday the day with lower incidents.

Furthermore, analyzing the months of the year, October, November, February, and April emerge as the most critical periods for illegal parking incidents which may be attributed to the holidays occurring during those months, as evidenced by the multi-row card highlighting the most critical holidays falling within these periods. In contrast, August exhibits the lowest number of illegalities, likely due to the absence of holidays during that month. Among the holidays, Carnival, All Saints' Day, and Restoration of Independence stand out as the most critical, while New Year's Day, Valentine's Day, and Christmas Day exhibit lower occurrences. Interestingly, although New Year's Day and Christmas Day are typically calm holidays, their respective Eves experience a significantly higher number of illegal parking incidents. Additionally, it is noteworthy that for holidays the hour period from 12 pm to 3 am is also identified as one of the most critical periods, contrasting with non-holiday periods.

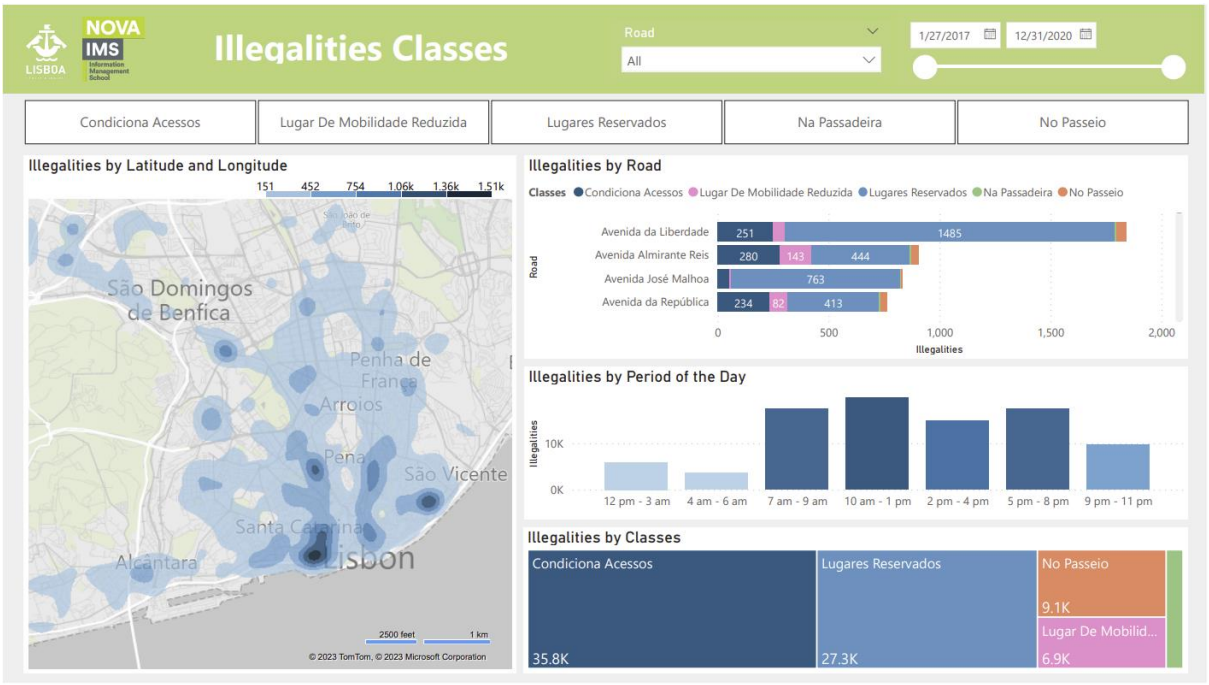


Figure 16 – Dashboard of Parking Illegalities Classes

The third and final page of the dashboard. represented in Figure 16, provides a detailed analysis of illegal parking occurrences categorized by classes. This page offers interactive features, including slicers by road, timeframe, and the most important slicer by classes where we can select only one or multiple classes to analyze. The main objective of this page is for users to always select a specific illegal class using the slicer, so they can explore the characteristics and patterns associated with that particular class in greater detail.

The Heat map identifies hotspots and areas with higher concentration of illegalities, the horizontal stacked bar chart illustrates the occurrences of illegalities by road, segmented by classes allowing for a comparative analysis of different roads and their associated classes of illegalities, and the vertical bar chart showcases the evolution of illegal parking occurrences throughout the day. Finally, the treemap visualization presents a hierarchical view of the illegality's classes with relative proportions and frequencies of all the different classes.

Through this page, it is observed that conditions access exhibits higher number of illegal parking occurrences across the Lisbon, with concentration of occurrences in central areas and those surrounding downtown, being the roads "Rua da Portela", "Avenida Almirante Reis" and "Avenida da Liberdade" the most critical ones and the period between 10 am to 1pm the one with the highest occurrences. In addition, regarding the Reserved class, the second most critical class, presents more occurrences in the downtown area of Lisbon, but the most critical streets are "Avenida da Liberdade" and "Avenida José Malhoa" with the periods 7am to 9pm and 10 am to 1pm the worst ones.

The disabled class shows higher occurrences of illegal parking in areas such as "Arroios," "Avenidas Novas," and near the "Estadio da Luz," with the period from 5 pm to 8 pm being particularly critical. Finally, for crosswalk and sidewalk, the heatmaps reveal that these classes are less spread around Lisbon, having as critical areas the ones close to "Feira da Ladra", in the case of sidewalk, and, in the case of crosswalk, in "Areeiro" and center of Lisbon and presenting the

5. CONCLUSIONS

The final chapter of this thesis presents a concise summary of the research conducted, highlighting the key findings and conclusions. Additionally, the chapter addresses the limitations encountered during the research process and provides recommendations for future work that can be explored to further enhance the understanding and application of the study's findings.

5.1. SUMMARY & IMPLICATIONS

Illegal parking is a persistent challenge in urban areas, causing traffic disruptions, safety concerns, and inconvenience for both drivers and pedestrians. These situations are very complex, so they require efficient and optimized support systems, such as Machine Learning models and business intelligence tools. In this research, our objective was to develop a predictive model capable of estimating the number of parking violations in Lisbon's most critical streets during different periods of the day. We explored various Machine Learning models, considering factors like road characteristics and air quality data, with the aim of finding a model that could generalize well to new data and produce accurate predictions. Among the models evaluated, the K-nearest neighbors model exhibited the most promising results, demonstrating lower errors and a reduced risk of overfitting. Lastly, we developed an interactive and user-friendly dashboard that enhances our understanding of illegal parking behavior, providing several valuable insights for Lisbon authorities.

Our findings reveal that areas in close proximity to downtown Lisbon experience a higher number of illegal parking occurrences, with "Avenida da Liberdade" identified as the most critical street, and the "conditions access" class the most prominent category. Furthermore, the analysis suggests that the time period between 10 am to 1 pm is particularly susceptible to infractions and Carnival exhibits the highest frequency of illegal parking incidents. Regarding the months, august is the most chill in terms of illegalities while the months of February, March, and October emerge as the most critical periods.

We envision these tools, the predictive model and interactive dashboards, to be a valuable resource for empowering responsible authorities by considering diverse categories of illegal parking and providing detailed spatial insights, enabling more targeted, efficient, and effective parking management policies. Through analysis of illegal parking patterns, authorities can optimize the allocation of human resources, and swiftly respond to unforeseen situations. Furthermore, this research offers the opportunity to identify critical areas that would benefit from the implementation of parking regulations and providing valuable information for the optimized deployment of police officers to effectively address illegal parking incidents.

5.2. LIMITATIONS & FUTURE WORK

During the course of this study, several limitations were identified that may have impacted the comprehensiveness and depth of the analysis. The major limitation of this project was the presence of an imbalanced dataset, which restricted our predictive modeling to focus exclusively on the top 10 critical roads grouped by hour periods. As a result, the dataset used for modeling was relatively small, potentially leading to biased predictions, increased variance, and reduced performance on unseen data. Moreover, when the data was grouped, a drawback emerged as the temporal order and

sequential patterns within each group were lost. In this way, future studies should consider expanding the dataset to include a more diverse range of roads and time periods, thus improving the representation of different scenarios, enabling more reliable predictions and capturing the patterns, trends, and seasonality that occur over time.

Additionally, this study is subject to limitations related to the influence of officer patrolling on the number of parking illegalities. It is important to acknowledge that areas with fewer officer presences and monitoring activities tend to exhibit a lower frequency of recorded illegal parking incidents which may introduce bias and affect the observed patterns and trends in the data. To address this limitation and enhance the analysis, future work should incorporate information on the levels of officer patrolling in the areas and periods of time studied. By considering the density and frequency of officer patrols as an additional feature, the predictive model can better account for the potential impact of law enforcement activities on the occurrence of parking illegalities. Furthermore, integrating data on officer schedules, shifts, and deployment strategies can provide valuable insights into the relationship between enforcement efforts and observed patterns of illegal parking.

Moreover, in future work, the incorporation of more updated data on illegal parking incidents will enhance the predictive model's accuracy by capturing recent trends and patterns, enabling more accurate predictions and timely interventions. In addition, this research could be expanded to another city like Porto where it would provide valuable insights into the generalizability and transferability of the findings. By considering a different urban context, the predictive model can be tested and validated in a new setting, enabling a broader understanding of the factors influencing illegal parking across multiple cities.

To conclude, another limitation of our study was the limited time range of the traffic data provided by Waze, which prevented us from including this potentially relevant variable in our predictive model, as doing so would have resulted in missing values and further reduction of our already small dataset. However, in future work, it is essential to address this limitation by considering the inclusion of traffic data with a more extensive time range since it would not only increase the predictive power of the model but also provide a more comprehensive understanding of the factors influencing parking violations in Lisbon. This, in turn, can contribute to more effective urban planning, traffic management, and enforcement strategies to mitigate illegal parking incidents and improve overall traffic flow and safety in the city.

BIBLIOGRAPHICAL REFERENCES

United Nations, 2018. World Urbanization Prospects. Demographic Research 12.

Apté, C., & Weiss, S. (1997). Data mining with decision trees and decision rules. *Future Generation Computer Systems*, *13*(2), 197–210. https://doi.org/10.1016/S0167-739X(97)00021-6

Bahrami, S., Vignon, D., Yin, Y., & Laberteaux, K. (2021). Parking management of automated vehicles in downtown areas. *Transportation Research Part C: Emerging Technologies*, *126*, 103001. https://doi.org/10.1016/j.trc.2021.103001

Basri Said, L., & Syafey, I. (2021). The scenario of reducing congestion and resolving parking issues in Makassar City, Indonesia. *Case Studies on Transport Policy*, *9*(4), 1849–1859. https://doi.org/10.1016/j.cstp.2021.10.004

Basu, R., & Ferreira, J. (2020). Understanding household vehicle ownership in Singapore through a comparison of econometric and machine learning models. *Transportation Research Procedia*, *48*, 1674–1693. https://doi.org/10.1016/j.trpro.2020.08.207

Boragno, S., Boghossian, B., Black, J., Makris, D., & Velastin, S. (2007). A DSP-based system for the detection of vehicles parked in prohibited areas. *2007 IEEE Conference on Advanced Video and Signal Based Surveillance*, 260–265. https://doi.org/10.1109/AVSS.2007.4425320

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step data mining guide*. https://www.semanticscholar.org/paper/CRISP-DM-1.0%3A-Step-by-step-data-mining-guide-Chapman-Clinton/54bad20bbc7938991bf34f86dde0babfbd2d5a72

Chen, W., & Yeo, C. K. (2019). Unauthorized Parking Detection using Deep Networks at Real Time. *2019 IEEE International Conference on Smart Computing (SMARTCOMP)*, 459–463. https://doi.org/10.1109/SMARTCOMP.2019.00088

Conway, A. J., Thuillier, O., Dornhelm, E., & Lownes, N. E. (2013). *Commercial Vehicle-Bicycle Conflicts: A Growing Urban Challenge* (No. 13–4299). Article 13–4299. Transportation Research Board 92nd Annual MeetingTransportation Research Board. https://trid.trb.org/view/1242509

Dietterich, T. G. (2000). Ensemble Methods in Machine Learning. *Multiple Classifier Systems*, 1–15. https://doi.org/10.1007/3-540-45014-9_1

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, *29*(5), 1189–1232.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, *33*(1), 1–22.

Fulman, N., Benenson, I., & Ben-Elia, E. (2020). Modeling parking search behavior in the city center: A game-based approach. *Transportation Research Part C: Emerging Technologies*, *120*, 102800. https://doi.org/10.1016/j.trc.2020.102800

Galatioto, F., & Bell, M. C. (2007). Simulation of illegal double parking: Quantifying the traffic and pollutant impacts. *Proc. 4th Int. SIIV Congr.*, 12–14.

Gao, J., & Ozbay, K. (2016, January 1). *Modeling Double Parking Impacts on Urban Street*.

Gao, J., & Ozbay, K. (2017, January 1). *A Data-driven Approach to Predict Double Parking Events Using Machine Learning Techniques*.

Gao, S., Li, M., Liang, Y., Marks, J., Kang, Y., & Li, M. (2019). Predicting the spatiotemporal legality of on-street parking using open data and machine learning. *Annals of GIS*, *25*(4), 299–312. https://doi.org/10.1080/19475683.2019.1679882

Group, G. M. (2019, June 4). *42 minutos de fila por dia: Lisboa é a cidade ibérica com mais trânsito*. https://www.dn.pt/dinheiro/42-minutos-de-fila-por-dia-lisboa-e-a-cidade-iberica-com-mais-transito-10976480.html

Harriet, T., & Poku, K. (2013). *An Assessment of Traffic Congestion and Its Effect on Productivity in Urban Ghana*. *4*(3).

Hennessy, D. A., & Wiesenthal, D. L. (1999). Traffic congestion, driver stress, and driver aggression. *Aggressive Behavior*, *25*(6), 409–423. https://doi.org/10.1002/(SICI)1098-2337(1999)25:6<409::AID-AB2>3.0.CO;2-0

IMT - Manutenção. (2018). https://www.imt-ip.pt/

Jardim, B., Alpalhão, N., Sarmento, P., & de Castro Neto, M. (2022). The Illegal Parking Score–Understanding and predicting the risk of parking illegalities in Lisbon based on spatiotemporal features. *Case Studies on Transport Policy*, *10*(3), 1816–1826.

Jardim, B., Castro Neto, M. de, Alpalhão, N., & Calçada, P. (2022). The daily urban dynamic indicator: Gauging the urban dynamic in Porto during the COVID-19 pandemic. *Sustainable Cities and Society*, *79*, 103714. https://doi.org/10.1016/j.scs.2022.103714

Jennath, H. S., Adarsh, S., Chandran, N. V., Ananthan, R., Sabir, A., & Asharaf, S. (2019). Parkchain: A Blockchain Powered Parking Solution for Smart Cities. *Frontiers in Blockchain*, *2*. https://www.frontiersin.org/articles/10.3389/fbloc.2019.00006

Jiang, J., Chen, Y.-C., & Hsieh, H.-P. (2020). Detection of Illegal Parking Events Using Spatial-Temporal Features. *Proceedings of the 28th International Conference on Advances in Geographic Information Systems*, 667–668. https://doi.org/10.1145/3397536.3428350

Kladeftiras, M., & Antoniou, C. (2013). Simulation-based assessment of double-parking impacts on traffic and environmental conditions. *Transportation Research Record*, *2390*(1), 121–130.

Kotb, A. O., Shen, Y., & Huang, Y. (2017). Smart Parking Guidance, Monitoring and Reservations: A Review. *IEEE Intelligent Transportation Systems Magazine*, *9*(2), 6–16. https://doi.org/10.1109/MITS.2017.2666586

Kumar, P., Nigam, S. P., & Kumar, N. (2014). Vehicular traffic noise modeling using artificial neural network approach. *Transportation Research Part C: Emerging Technologies*, *40*, 111–122. https://doi.org/10.1016/j.trc.2014.01.006

Lin, B., & Du, Z. (2015). How China's urbanization impacts transport energy consumption in the face of income disparity. *Renewable and Sustainable Energy Reviews*, *52*, 1693–1701. https://doi.org/10.1016/j.rser.2015.08.006

Liu, Y., Wang, W., Ding, C., Guo, H., Guo, W., Yao, L., Xiong, H., & Tan, H. (2012). Metropolis Parking Problems and Management Planning Solutions for Traffic Operation Effectiveness. *Mathematical Problems in Engineering*, *2012*, 1–6. https://doi.org/10.1155/2012/678952

Marsden, G. (2006). The evidence base for parking policies—A review. *Transport Policy*, *13*(6), 447–457. https://doi.org/10.1016/j.tranpol.2006.05.009

Minka, T. (2000). *Bayesian linear regression*. Citeseer.

Morillo, C., & Campos, J. M. (2014). On-street Illegal Parking Costs in Urban Areas. *Procedia - Social and Behavioral Sciences*, *160*, 342–351. https://doi.org/10.1016/j.sbspro.2014.12.146

National Smart City Strategy. (2022, June 20). *Portugal Digital*. https://portugaldigital.gov.pt/en/promote-more-digital-public-services/more-digital-territories/national-smart-city-strategy/

Ng, C. K., Cheong, S. N., Yap, W. W.-J., & Foo, Y. L. (2018). Outdoor Illegal Parking Detection System Using Convolutional Neural Network on Raspberry Pi. *International Journal of Engineering & Technology*, *7*(3.7), Article 3.7. https://doi.org/10.14419/ijet.v7i3.7.16197

Nguyen, S., Salcic, Z., & Zhang, X. (2018). Big Data Processing in Fog—Smart Parking Case Study. *2018 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Ubiquitous Computing & Communications, Big Data & Cloud Computing, Social Computing & Networking, Sustainable Computing & Communications (ISPA/IUCC/BDCloud/SocialCom/SustainCom)*, 127–134. https://doi.org/10.1109/BDCloud.2018.00031

Nourinejad, M., Gandomi, A., & Roorda, M. J. (2020). Illegal parking and optimal enforcement policies with search friction. *Transportation Research Part E: Logistics and Transportation Review*, *141*, 102026. https://doi.org/10.1016/j.tre.2020.102026

Parmar, J., Das, P., & Dave, S. M. (2020). Study on demand and characteristics of parking system in urban areas: A review. *Journal of Traffic and Transportation Engineering (English Edition)*, *7*(1), 111–124. https://doi.org/10.1016/j.jtte.2019.09.003

Peterson, L. E. (2009). K-nearest neighbor. *Scholarpedia*, *4*(2), 1883. https://doi.org/10.4249/scholarpedia.1883

Porikli, F. (2007). Detection of temporarily static regions by processing video at different frame rates. *2007 IEEE Conference on Advanced Video and Signal Based Surveillance*, 236–241. https://doi.org/10.1109/AVSS.2007.4425316

Ritchie, H., & Roser, M. (2018). Urbanization. *Our World in Data*. https://ourworldindata.org/urbanization

Shoup, D. C. (2006). Cruising for parking. *Transport Policy*, *13*(6), 479–486. https://doi.org/10.1016/j.tranpol.2006.05.005

Šiurytė, A., & Davidavičienė, V. (2016, February 11). *AN ANALYSIS OF KEY FACTORS IN DEVELOPING A SMART CITY*. https://doi.org/10.3846/mla.2015.900

Spiliopoulou, C., & Antoniou, C. (2012). Analysis of Illegal Parking Behavior in Greece. *Procedia - Social and Behavioral Sciences*, *48*, 1622–1631. https://doi.org/10.1016/j.sbspro.2012.06.1137

The road to seamless mobility | McKinsey. (2019). https://www.mckinsey.com/capabilities/sustainability/our-insights/the-road-to-seamless-urban-mobility

Transforming our World: The 2030 Agenda for Sustainable Development. (2015). United Nations Population Fund. https://www.unfpa.org/resources/transforming-our-world-2030-agenda-sustainable-development

United Nations. (2019). *Standards for the Sustainable Development Goals*. UN. https://doi.org/10.18356/6a0e015b-en

Using Machine Learning to Predict Parking Difficulty. (2017, February 3). https://ai.googleblog.com/2017/02/using-machine-learning-to-predict.html

Vehicle sales mirror economic growth (2008-2021 trend). (2020, January 1). *ACEA - European Automobile Manufacturers' Association*. https://www.acea.auto/figure/vehicle-sales-mirror-economic-growth-2008-2021-trend/

Vlahov, D. (2002). Urbanization, Urbanicity, and Health. *Journal of Urban Health: Bulletin of the New York Academy of Medicine*, *79*(90001), 1S – 12. https://doi.org/10.1093/jurban/79.suppl_1.S1

Wang, S.-C. (2003). Artificial Neural Network. In S.-C. Wang (Ed.), *Interdisciplinary Computing in Java Programming* (pp. 81–100). Springer US. https://doi.org/10.1007/978-1-4615-0377-4_5

Weinberger, R. R., Millard-Ball, A., & Hampshire, R. C. (2020). Parking search caused congestion: Where's all the fuss? *Transportation Research Part C: Emerging Technologies*, *120*, 102781. https://doi.org/10.1016/j.trc.2020.102781

Weisberg, S. (2005). *Applied Linear Regression*. John Wiley & Sons.

Weisbrod, G., Vary, D., & Treyz, G. (2003). Measuring Economic Costs of Urban Traffic Congestion to Business. *Transportation Research Record*, *1839*(1), 98–106. https://doi.org/10.3141/1839-10

World Cities Report 2022. (2022). https://unhabitat.org/wcr/

Yin, Z., Xiong, H., Zhou, X., Goldberg, D. W., Bennett, D., & Zhang, C. (2019). A Deep Learning based Illegal Parking Detection Platform. *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, 32–35. https://doi.org/10.1145/3356471.3365233

Zheng, Y., Rajasegarar, S., & Leckie, C. (2015). Parking availability prediction for sensor-enabled car parks in smart cities. *2015 IEEE Tenth International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP)*, 1–6. https://doi.org/10.1109/ISSNIP.2015.7106902

Zoika, S., Tzouras, P. G., Tsigdinos, S., & Kepaptsoglou, K. (2021). Causal analysis of illegal parking in urban roads: The case of Greece. *Case Studies on Transport Policy*, *9*(3), 1084–1096. https://doi.org/10.1016/j.cstp.2021.05.009