UNIVERSIDADE FEDERAL DE ITAJUBÁ - UNIFEI
MASTER DEGREE PROGRAM IN
COMPUTER SCIENCE AND TECHNOLOGY

Driver's Behavior Classification in Vehicular Communication Networks for Commercial Vehicles.

Lucas Gomes de Almeida

Itajubá, June 5, 2023

UNIVERSIDADE FEDERAL DE ITAJUBÁ - UNIFEI
MASTER DEGREE PROGRAM IN
COMPUTER SCIENCE AND TECHNOLOGY

Lucas Gomes de Almeida

Driver's Behavior Classification in Vehicular Communication Networks for Commercial Vehicles.

Dissertation submitted to Masters Degree Program in Computer Science and Technology as part of requirements to obtain the Master (M.Sc.) in Computer Science and Technology title.

Concentration Area: Computing Mathematics

Orientador: Prof. Dr. Bruno Tardiole Kuehne
Coorientador: Prof. Dr. Otávio de Souza Martins Gomes

June 5, 2023
Itajubá

Lucas Gomes de Almeida

Driver's Behavior Classification in Vehicular Communication Networks for Commercial Vehicles/ Lucas Gomes de Almeida. – Itajubá, June 5, 2023-

91 p. : il. (algumas color.) ; 30 cm.

Orientador: Prof. Dr. Bruno Tardiole Kuehne

Master Thesis'

Universidade Federal de Itajubá - UNIFEI

Master Degree Program in Computer Science and Technology, June 5, 2023.

1. Driver Behavior. 2. Clustering. I. Orientador Professor Bruno Tardiole Kuehne. II. Universidade Federal de Itajubá - UNIFEI. III. Faculdade de xxx. IV. Título

CDU 07:181:009.3

Agradecimentos

First and foremost, I would like to express my gratitude to the Almighty God for bestowing upon me countless blessings, knowledge, and opportunities, which have enabled me to accomplish this thesis.

I would also like to extend my deepest appreciation to my esteemed professors, Bruno Tardiole and Otavio Gomes, for their invaluable patience and feedback, which have been instrumental in shaping my research. I am equally grateful to my professors, Stephan and Ovidiu, from Derby University, who generously shared their knowledge and expertise with me.

I must also acknowledge the unwavering support and encouragement that I received from my classmates and cohort members, especially Viviane Cordeiro, who played a significant role in supporting me during this thesis.

Furthermore, I am grateful to Volkswagen Truck and Bus for their generous support and encouragement of my research, without which this endeavor would not have been possible.

Lastly, I would like to express my heartfelt appreciation to my family, especially my parents and wife Luciana, for their unwavering belief in me, which has been a constant source of motivation and inspiration throughout this process.

"No man is better than a machine, and no machine is better than a man with a machine." Richard Bookstaber

Abstract

Vehicles are becoming more intelligent and connected due to the demand for faster, efficient, and safer transportation. For this transformation, it was necessary to increase the amount of data transferred between electronic modules in the vehicular network since it is vital for an intelligent system's decision-making process. Hundreds of messages travel all the time in a vehicle, creating opportunities for analysis and development of new functions to assist the driver's decision. Given this scenario, the dissertation presents the results of research to characterize driving styles of drivers using available information in vehicular communication network.

This master thesis focuses on the process of information extraction from a vehicular network, analysis of the extracted features, and driver classification based on the extracted data. The study aims to identify aggressive driving behavior using real-world data collected from five different trucks running for a period of three months. The driver scoring method used in this study dynamically identifies aggressive driving behavior during predefined time windows by calculating jerk derived from the acquired data. In addition, the K-Means clustering technique was explored to group different behaviors into data clusters.

Chapter 2 provides a comprehensive overview of the theoretical framework necessary for the successful development of this thesis. Chapter 3 details the process of data extraction from real and uncontrolled environments, including the steps taken to extract and refine the data. Chapter 4 focuses on the study of features extracted from the preprocessed data, and Chapter 5 presents two methods for identifying or grouping the data into clusters.

The results obtained from this study have advanced the state-of-the-art of driver behavior classification and have proven to be satisfactory. The thesis addresses the gap in the literature by using data from real and uncontrolled environments, which required preprocessing before analysis. Furthermore, the study represents one of the pioneering studies conducted on commercial vehicles in an uncontrolled environment.

In conclusion, this thesis provides insights into the development of driver behavior classification models using real-world data. Future research can build upon the techniques presented in this study and further refine the classification models. The thesis also addresses the threats to validity that were mitigated and provides recommendations for future research.

Key-words: Driver Behavior. Vehicle. Controller Area Network. K-Means. Clustering. Data Analysis.

List of Figures

List of Tables

ABS - Antilock Brake System

AI - Artificial Intelligence

CAN - Controller Area Network

CEO - Chief of Executive Office

CNT - National Transport Confederation

CRC - Cyclic Redundancy Check

DLC - Data Length Code

DTC - Diagnostic Trouble Code

DAQ - Data Acquisition System

DBSCAN - Density-based spatial clustering of applications with noise

ECU - Electronic Control Unit

EOF - End of Frame

IoT - Internet of Things

IoV - Internet of Vehicles

LLC - Logic Link Control

MAC - Medium Access Control

PCA - Principal Component Analysis

PGN - Parameter Group Number

RQ - Research Question

SAE - Society of Automotive Engineers

SOF - Start of Frame

t-SNE - t-Stochastic Neighbour Embedding

UNIFEI - Federal University of Itajubá

VWCO - Volkswagen Truck and Bus

WHO - World Health Organization

Contents

1 Introduction

Technological inventions have profoundly impacted human history, such as the creation of the automobile. Its influence is mainly due to the unparalleled ease and mobility it provides for society. This creation impacted the economy, industry, and even American culture. More than 4.2 million people worldwide work directly in the auto industry, generating over 2 trillion dollars in the economy every year [7]. On the flip side, with 1.2 billion vehicles traveling worldwide, accidents happen daily. To bring more safety and efficiency to vehicles, the development of vehicle electronics has gained significant interest in recent years.

As the automotive industry transforms, the world is experiencing the age of the Internet of Things (IoT), where each object can become intelligent through sensing and its connection to the internet. Although using smartwatches, smart TVs, and other household items is one of the main applications of IoT, vehicles also play a vital role in this revolution. In the near future, vehicles will be interconnected throughout the logistics chain, consisting of the automaker, dealership, insurance company, customer, and highways.

Over the last few decades, the automotive industry has witnessed a surge in the utilization of electronic modules to facilitate the integration of new technologies in a connected world. According to a study by Fugiglando in 2019, a modern vehicle now contains around 70 electronic control units (ECU) [8], which is a stark contrast to the industry's beginning when there were none. Moreover, most contemporary cars come equipped with over 400 sensors that capture various information crucial for the vehicle's efficient functioning.

Recognizing the industry's future lies in the software behind the vehicle rather than the vehicle itself, automakers worldwide have shifted their focus from selling cars to selling software [9]. Consequently, it is critical to extract relevant data from the vehicle without incurring additional costs to the customer to stay ahead in an increasingly competitive world.

One of the many types of information that can be obtained through data in a vehicular network is the behavior of the driver in traffic. This information is crucial in assessing the safety of traffic, as an aggressive driver can put many lives at risk. The focus of this master's thesis is to develop an algorithm that can determine whether a driver is exhibiting aggressive behavior in traffic or not. The algorithm will be based on existing information in the vehicular communication network, and the aim is to deepen our understanding of how to accurately identify and differentiate aggressive driving behavior from safe driving behavior.

The study will identify and enumerate the differences between safe and aggressive driving behaviors to gain a complete understanding of a driver's behavior. While there have been studies on safe and unsafe behavior, they are limited in number and often conducted in controlled environments and without commercial vehicles, which pose a greater potential risk for traffic compared to passenger vehicles.

This work focuses on studying commercial vehicles due to the lack of research in this area compared to passenger vehicles. While both types of vehicles share similarities, it is important to consider the differences in commercial vehicles. For instance, commercial drivers are professionals who are trained and paid for their work. They are also responsible for transporting cargo and must exercise caution in doing so. However, concerns over fuel consumption and delivery time can sometimes lead to aggressive driving behavior in commercial drivers. Additionally, the amount of time spent behind the wheel can also impact a driver's performance.

To gain insight into the current situation in Brazil regarding traffic accidents involving commercial vehicles, a survey was conducted in 2018 by the National Transport Confederation (CNT) [1]. This survey highlighted the number of accidents involving these vehicles. Figure 1, extracted from the aforementioned survey, illustrates that, despite recent years' decrease in the number of accidents, the number remains high and pertinent to the Brazilian context. The blue line represents the number of accidents involving commercial vehicles without casualties, while the red line represents the number of accidents with casualties. Notably, the red line exhibits greater stability than the blue line, underscoring the prevalence of fatal accidents in Brazil.

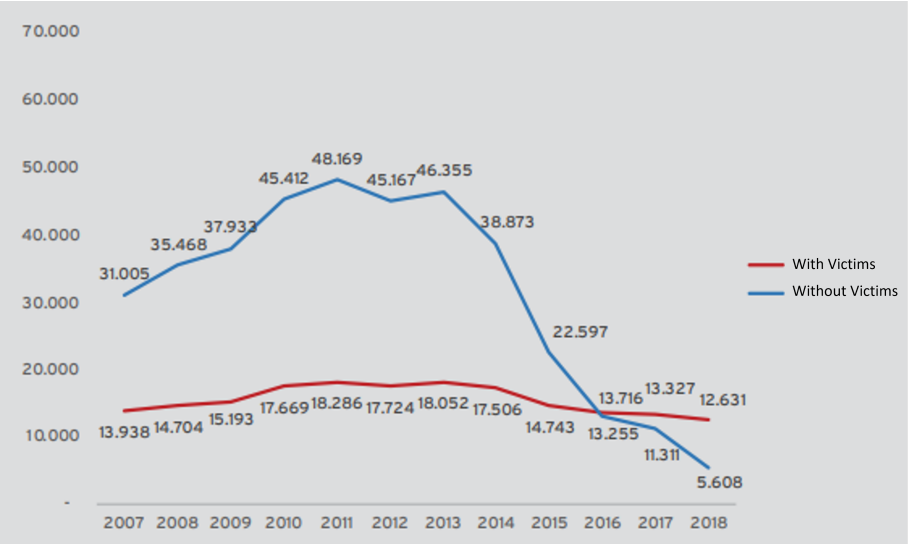

Figure 1. Number of accidents with casualties occurring on federal highways by type of outcome – Brazil – 2007 to 2018. Extracted from [1]

The aforementioned survey also provided insights into the distribution and leading causes of accidents. It can be observed that human factors, such as inattention or improper

driver behavior, are responsible for the majority of accidents. Furthermore, collisions, particularly resulting from speeding on highways, are the most prevalent type of accident. These data from 2007 to 2018 are illustrated in Figures 2 and 3. Figure 2 shows that collisions account for 67.4% of the casualties, making it the most common cause of death in traffic accidents. This type of accident is often linked to aggressive driving behavior, highlighting the potential of this research. On the other hand, Figure 3 indicates that human factors are the leading cause of accidents, with a lack of attention being responsible for 29% of them.

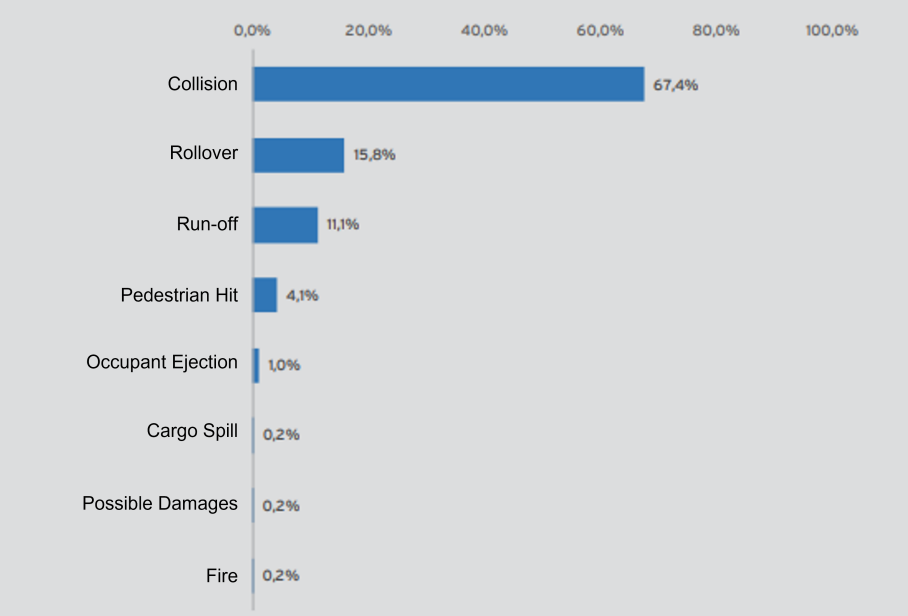

Figure 2. Distribution of deaths in accidents with victims involving truck racing on federal highways by type of accident – Brazil – accumulated 2007 to 2018. Extracted from [1]

Determining drivers' aggressive behavior can provide valuable insights into their participation in traffic and help to reduce the societal costs of accidents. The information generated by this research can also be leveraged by insurance companies and governments to improve road safety. In fact, some governments, such as Belgium and the Netherlands, have already initiated campaigns to address this issue. Globally, speeding on highways remains one of the primary causes of accidents, and this issue could be effectively addressed through the implementation of a data analysis algorithm capable of detecting unsafe patterns of driver behavior. Data analysis techniques are crucial for analyzing the vast and complex data sets generated by vehicular networks.

Recent studies in this field have focused on distinguishing between aggressive and expected driver behavior. Halim, Kalsoom, and Baig [10] conducted a study demonstrating that driver behavior can be determined through the integration of data collected from participants in a driving simulator. The variables considered in this analysis were the

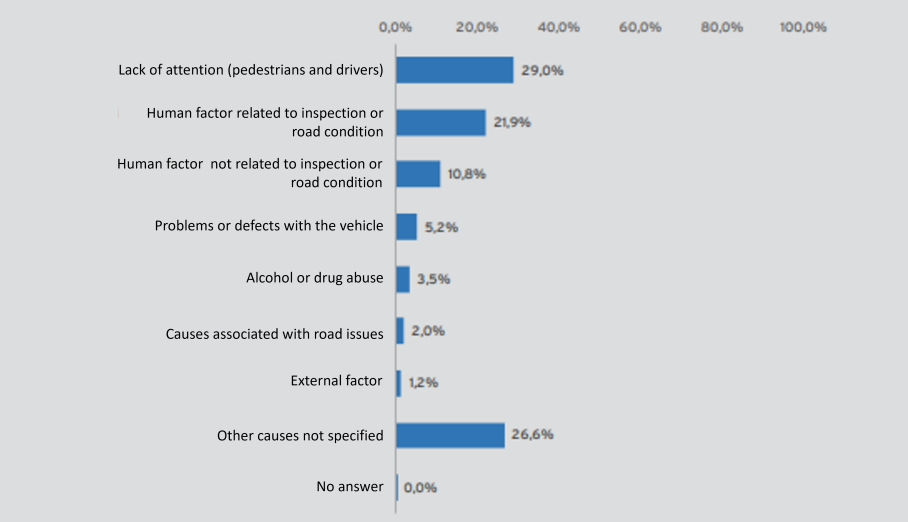

Figure 3. Distribution of accidents with victims involving a truck that occurred on federal highways because of the accident – Brazil – accumulated 2007 to 2018. Extracted from [1]

maximum speed, acceleration, use of the brake pedal, and frequency of horn usage in the vehicle.

Driving simulators have been extensively used in various research studies to analyze data on driver behavior in vehicular communication networks, thereby significantly increasing knowledge in this field [11]. These simulators offer several advantages, including lower validation costs than actual vehicles, training without posing any real risk to the driver, and the ability to maintain the test in the same condition for different drivers. However, there are also certain drawbacks associated with these simulators, the most significant being that the results obtained are different from those in the real world, as found in the research by Godley et al [12]. Although it is possible to reproduce a simulator with a high degree of fidelity, the article indicates that the results obtained only provide an indication of the direction of human behavior in a given activity, but not the exact magnitude of this event in real life.

This thesis presents a novel approach to determine the driver's profile. The proposed methodology involves the collection of precise data from diverse drivers, which will be subsequently refined to extract a profile that can be utilized by a vehicle fleet.

1.1 Research Context

This study was conducted through a partnership between Volkswagen Truck and Buses, hereafter referred to as VWCO, and the Federal University of Itajubá - UNIFEI. VWCO is one of the world's largest manufacturers of commercial vehicles, and was the best-selling brand in Brazil in 2021. The automaker produces a range of vehicles, from

light models such as the Delivery Express, to extra-heavy models such as the Meteor. The company's headquarters are located in Resende, in the interior of Rio de Janeiro state, where all VWCO national vehicles are manufactured. VWCO has been responsible for significant innovations in the Brazilian market, such as e-Delivery, the brand's first electric truck. Safety has always been a critical concern for the company, and thus implementing an algorithm capable of identifying aggressive driving behavior is of utmost interest. Furthermore, analyzing driver behavior through the existing data in the vehicular network presents a significant academic challenge and a powerful resource for future applications in the vehicular context. This chapter, along with section 1.2, will discuss the motivations and implications of this information for the automotive industry. Figure 4 displays the current models in the company's lineup.

Figure 4. Volkswagen Truck and Bus products. [2]

1.2 Motivation

The increasing demand for low-cost technology integration necessitates that vehicular communication networks accommodate a larger volume of information, offer higher speeds, and interconnect smart devices while maintaining their reliability.

The main objective of this thesis is to explore a function that has numerous applications in the automotive field, which is the identification and classification of the driver's profile. It is possible to observe differences in the way people use their vehicles on a daily basis while driving in traffic. Some drivers use the accelerator pedal more abruptly, while others apply the brake more smoothly, which allows us to identify a unique driver fingerprint. As Ezzini et al. [13] pointed out, the concept of a driver fingerprint obtained from analyzing existing data on the CAN network is not new, as many studies using machine learning techniques have demonstrated its feasibility. However, there are only a

few studies conducted under natural application conditions, particularly for commercial vehicles.

This work aims to complete the entire process of identifying the driver's profile by extracting data from the vehicular network, analyzing the relevance of the extracted data for composing this analysis, and ultimately determining the aggressiveness of the driver's behavior.

In recent years, both the scientific community and industry professionals have been engaged in developing the next generation of smart vehicles. This drive is largely motivated by the latest legislation implemented in Europe, particularly laws pertaining to active safety and autonomous vehicles. Such regulations have a global impact, including in Brazil where they may become law after an average of five years.

The author identified some possible applications of this research after identifying driver behavior, especially aggressive behavior:

1. **Assist in detecting another aggressive vehicle on the highway by developing an intelligent vehicle algorithm.** In recent years, the subject of intelligent and autonomous vehicles has gained significant attention. Before delving deeper into this topic, it is essential to establish a clear definition of what constitutes an intelligent vehicle. As stated by [5], an intelligent vehicle is a system that observes its driving environment, detects participants and pertinent information about the driving scene, and interprets these observations with a degree of understanding, enabling it to make predictions about future occurrences. Various technological advancements in the field of autonomous or semi-autonomous vehicles and driver assistance systems, including functions such as adaptive cruise control, pedestrian detection, and blind spot detection, are being developed by research groups and companies alike.

   One company that has been developing a fleet of autonomous vehicles for several years is Google [14]. They have presented a vehicle that integrates sensory abilities and allows control through sensors and actuators in vital parts of the vehicle, such as the gas pedal, steering, and sensors. The company has released several videos and demonstrated the vehicle's abilities. Another company that has made significant progress in this field, including the launch of vehicles, is Tesla. Elon Musk, the CEO and leader of the company, has conducted a series of interviews and press releases, highlighting all the advancements made by the company, which is considered one of the pioneers in automotive innovation today.

   To begin with, it is imperative to comprehend the concept of safe driving and the decisions required to create a successful autonomous vehicle. This study aims to contribute to this field, as mentioned previously, by introducing an algorithm capable of detecting aggressive driving behavior. This information can potentially

aid the decision-making process of an autonomous vehicle, making it a valuable asset in the future implementation of the Internet of Vehicles (IoV). The IoV is a concept where all vehicles in the network are interconnected, exchanging valuable information amongst each other. By advancing autonomous vehicles for society, we can eliminate one of the most significant causes of traffic accidents, which is human error. The identification of driver behavior leads to the second application topic for this research.

2. **Increase driving safety through accident prevention.** Each year, a significant number of accidents occur on highways worldwide. According to studies conducted by the World Health Organization (WHO), traffic-related accidents cause the death of 1.2 million people annually, with between 20 and 50 million individuals sustaining severe injuries. Adopting a purely economic perspective, these accidents result in the destruction of infrastructure and the depletion of resources. WHO's report further reveals that traffic accidents cost 2.5 trillion reais per year [5]. To provide a comparison, Brazil invests 129 billion reais in education annually, which is nearly 20 times the investment in all education in the country [1]. The human impact of these accidents is equally significant, with the loss of life being immeasurable for families and leading to long-lasting trauma.

   Researchers and automakers worldwide have been actively seeking solutions to reduce the alarming number of accidents on the world's highways. While technologies such as airbags and ABS brakes have significantly helped in reducing these numbers, they have largely been introduced to minimize the impact of accidents rather than prevent them. The next step in this field is to focus on preventing accidents from happening altogether. This can only be achieved by developing a system that can detect potentially hazardous conditions, determine the likelihood of an accident, and take appropriate measures to prevent it from occurring.

1.3 Objectives

In order to help the understanding of the entire process for identifying driver behavior, this thesis has three main goals and objectives:

1. The primary objective is to extract data from a vehicular communication network and group it to identify the variables and features present on the network.

2. Study these variables and reduce the dimensions of the data obtained in the experiment

3. Propose the implementation of data analysis and statistical techniques combined with time windows and data from the first two objectives to efficiently identify whether the driver fits into an aggressive driving profile.

1.4 Research Questions

The following questions will be addressed and answered throughout this dissertation:

- **RQ1:** What clustering algorithms should be used to cluster aggressive driver behaviors? **Motivation:** The reason behind choosing RQ1 is to acquire knowledge about clustering algorithms. It is also necessary to understand which algorithms could be applied to the acquired data since they came from actual vehicles.
- **RQ2:** What is the best time window to determine if the driver was aggressive? **Motivation:** The motivation behind RQ2 is to find the solution that delivers the most confidence interval and reliability to the established problem.

1.5 Dissertation Structure

This dissertation is organized as follows:

Chapter 2 provides an overview of the current state-of-the-art in vehicular communication networks, driver behavior analysis, and unsupervised learning techniques. The chapter introduces the concept of a Controller Area Network (CAN) and delves into the techniques used for data extraction and analysis. Chapter 3 explains in detail how data was extracted from the vehicular network and how it was processed prior to analysis. Chapter 4 outlines the methodology used to determine the features that were used in this research. Chapter 5 presents an actual case study, where data from six vehicles operating in various uncontrolled environments is analyzed, and drivers are classified into different behaviors and grouped into patterns. Finally, in chapter 6, the main results of this thesis are summarized, the threats to validity are identified, and recommendations for future work are provided.

2 Theorical Revision

This chapter provides an overview of the theoretical concepts essential for comprehending the development of this dissertation. Firstly, a concise overview of vehicular communication networks and the widely adopted CAN protocol in the automotive industry is presented. Subsequently, the current state-of-the-art studies on driver behavior are examined, focusing on identifying aggressive behavior in traffic. Lastly, data analysis techniques are explored to determine the most suitable approach to be applied in this thesis.

2.1 Data communication network

A data communication network serves as the standard communication environment between electronic devices. To determine the optimal network for a given application, several characteristics can define the way it operates, such as transmission rate and type of transmission, physical layer, and data encryption. The automotive industry has incorporated technological advances from computing and electronics since competitors began implementing electronic controls in vehicles. Consequently, vehicle controls have become more complex, distributed, and interconnected through embedded data communication networks.

In the context of connected devices, information is exchanged through formatted messages. A message typically consists of two primary fields: overhead, which pertains to communication control data, and payload, which represents the actual information being transmitted. The formatting and specific characteristics of these messages depend on the communication protocol being used for a given application.

2.1.1 Controller Area Network - CAN Bus

The implementation of vehicular communication networks between electronic modules began in the 1980s with the creation of the Controller Area Network (CAN) by Robert Bosch Gmbh. The primary objective was to optimize communication between different electronic modules by multiplexing information on the bus, reducing cabling between devices.

Initially, research focused on developing a low-speed network for non-critical applications. However, due to its significant acceptance by the automotive industry, new specifications were published, defining higher speeds for critical applications. This accep-

tance was driven by the relevance of functions and parameters for which the concept of distributed control could be applied.

In the 1990s, the International Organization for Standardization published the ISO 11898 standard based on work developed by Bosch, making CAN the standard for implementation in automotive applications.

The CAN Network is a serial communication protocol that efficiently supports real-time distributed controls with a high level of security.

The CAN network finds its application in various industrial sectors, from manufacturing processes, where it is implemented in production lines of manufacturing industries, to the manufactured goods sector, where it is present in refrigerators, satellites, boats, and automotive vehicles in general.

SAE has standardized specific protocols for the various vehicle segments in the automotive industry, where the broadest range of CAN network applications is found. For example, the SAE J1850 protocol is used for automobiles, while the ISO 11783 standard is the most commonly used for agricultural machinery. SAE J1939 is the most commonly used standard for commercial vehicles.

To standardize communication between devices via data communication networks, the International Organization for Standardization (ISO) defined a model in which communication is distributed across layers, with specific functions assigned to each layer and a common goal of organizing, prioritizing, and controlling the sending and receiving of messages. These layers, based on the OSI model and explained in ISO 11898, are:

1. **Application:** definition of user interface.
2. **Presentation:** responsible for data format and cryptography.
3. **Section:** responsible for remote link.
4. **Transport:** responsible for error handling.
5. **Network:** responsible for message address.
6. **Link:** responsible for data transmission format.
7. **Physical:** establish the main characteristics of the physical network environment.

The control information generated in the respective layers is added to the data to be transmitted in an encapsulation process.

The CAN network was designed with a simpler structure, where the message has less overhead and a larger data field. This optimization of the encapsulation process

ensures that the CAN message contains only the necessary information to control the communication, in addition to the physical data.

To access the required information on the network, it is essential to first study the data link layer of the CAN network. This layer is divided into two sublayers: the Logic Link Control (LLC) and the Medium Access Control (MAC).

The LLC sub-layer is the uppermost layer of the data link layer. It is responsible for various tasks, including message acceptance filtering, overload notification, message recovery management in the event of an error, and executing the link between the application layer and the MAC sub-layer through specific primitives [3].

The application layer software sends a primitive to the LLC layer, which includes the identifier information, data field size, and the data itself. The LLC sub-layer then sorts this information and passes it on to the MAC sub-layer for transmission.

The LLC frame contains three fields, as mentioned earlier:

- **Identifier field:** This field could be 11-bit or 29-bit for extended frames. It is the message identifier and its priority holder.
- **DLC field:** Data Length Code, that is, the size of the data field, in bytes. Four bits are used.
- **LLC data field:** It is the data set to be transmitted. It is used from 0 to 64 bits, depending on the information.

During the reception process, the LLC layer receives the information from the MAC layer, including the identifier. The LLC layer verifies the identifier and, if necessary, sends the LLC frame to the application layer to receive the message.

The MAC sub-layer is responsible for encapsulating and de-encapsulating messages, error management, and accessing the bus. It also performs the serialization or deserialization of the frame to transmit or receive the message. This sub-layer is considered the lowest sub-layer of the data link layer. The MAC frame contains the following information, which is visually detailed in Figure 5 and explained in each bullet point:

Figure 5. Visual representation of a CAN frame. Extracted from [3]

- **Start of Frame (SOF):** It is the dominant bit that determines the beginning of the message.
- **Arbitration Field:** Contains part of the information from the LLC layer, including a field that determines the priority and access to the bus. This field is 11 bits in size for standard frames and 29 bits for extended frames.
- **Control Field:** Indicates the number of bytes of the data field. It is called data length code (DLC) and contains four bits.
- **Data Field:** Data field, which can contain up to 64 bits.
- **Cyclic Redundancy Check Field (CRC):** It is one of the tools to detect if the message is correct and contains 16 bits.
- **Acknowledge field:** Field destined to confirm the receipt of a message by the nodes and contains 2 bits.
- **End of Frame (EOF):** Contains six consecutive recessive bits determining the end of the message.

The MAC sub-layer communicates with the LLC sub-layer through primitives to build a CAN frame.

2.1.2 Protocol SAE J1939

As mentioned in section 2.1.1, the SAE J1939 protocol was developed by the industry for use in commercial vehicles. This open standard defines the parameters required for technologies applied to commercial vehicles, as well as messages, identifiers, priorities, and parameter grouping with a common characteristic, all aimed at optimizing the use and transmission of information.

In the context of this protocol, access to the CAN bus is determined by the priority assigned to each message [32]. This priority is defined by the value of the message's identifier. Moreover, the CAN network is non-preemptive, meaning that a message's transmission is not interrupted by a higher priority message that becomes available for transmission during its transmission. Figure 6 illustrates the division of the CAN frame using the J1939 protocol. The frame consists of three main parts. The first part is the message priority, which determines the importance of the message in the network. This prioritization ensures that safety and legislation-related messages are given higher priority, allowing them to be received by other control units promptly. The second part is the message body, where all the generated and available data information is encapsulated. This section contains the actual payload of the message, carrying the relevant data that

needs to be transmitted. The last part of the frame is the source address, which serves the purpose of identifying the specific control unit responsible for sending the message. This address helps in establishing the origin of the message and aids in proper routing and communication within the network.

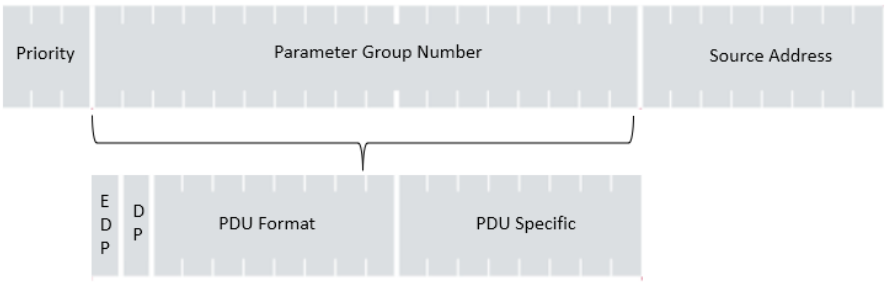

Figure 6. Visual representation of a CAN frame using Protocol J1939. Adapted from [4]

The SAE J1939 protocol has been published in sixteen volumes. The numbering of the volumes follows the sequence of the protocol name, which is not sequential. The main volumes of the protocol that are relevant to this master's thesis are:

- **SAE J1939/11 and SAE J1939/15:** refer to the physical layer of the CAN network. The protocol defines the characteristics of this layer and the required physical environment. Specifically, it mandates the use of a twisted pair with a ground mesh, as well as two channels: CAN High and CAN Low. The voltage levels and impedance matching resistors must also conform to the standard. The bus speed can be as high as 1 Mb/s, although the typical speed for commercial vehicles is 250 kb/s.
- **SAE J1939/21:** The data link layer is characterized in this volume, specifically as the exclusive use of extended frames with 29 bits. The frames are pre-defined, and the physical parameters form the data field of each message. Each parameter group is assigned a unique number called PGN (Parameter Group Number). The PGN is directly linked to the priority definition of each message, as it is part of the identifier. The lower the numerical value of the PGN, the higher the message's priority, and therefore the greater its chance of gaining access to the bus.
- **SAE J1939/31:** The network layer of the SAE J1939 protocol defines the components and their respective rules, parameters, and functions. These components are used to build a network system that allows for communication between different electronic control units (ECUs) in a vehicle. The components mentioned in this layer include bridges, routers, gateways, and repeaters, among others.
- **SAE J1939/71:** The Vehicle Application Layer volume defines all the significant parameters for a network application. For each parameter, it specifies the physical variable, resolution, range, measurement scale, the size of the word that will identify

it, and the associated parameter group. This volume plays a crucial role in the data extraction stage of this dissertation since it is necessary to understand the information that will be transmitted on the network.

- **SAE J1939/73:** Application Layer - Diagnostics. This volume defines the diagnostic protocol, which enables the system to detect abnormal situations, report them to the controller, and take appropriate action. To standardize the information, a unique Diagnostic Trouble Code (DTC) was created for each type of failure. The DTC allows for consistent and precise identification of the issue, facilitating efficient and effective troubleshooting.
- **SAE J1939/81:** Network Management. This volume defines the configuration of network addresses and the identification of each electronic module on the network. All modules are assigned a unique network address, which allows for easy identification of the message's origin and destination.

2.2 Driver Behavior Analysis

Defining a proper behavior for drivers is a complex task as there is no clear definition of what behavior entails. One of the primary objectives of this research is to establish a methodology to quantify driver aggression.

2.2.1 What is a behavior?

Intelligent vehicles represent a new frontier in the automotive industry, made possible only through the exponential advancement of technology, sensor implementation, and increased computing power in embedded modules. Despite this progress, discussions regarding aggressive behavior towards drivers are still relatively recent. There is no consensus among the scientific community on what constitutes a behavior, how to formulate it, and how to quantify it [5]. Human beings are one of the most complex machines ever created, making the definition of driving behavior a challenging task. Researchers often define driving behavior based on the specific scientific question they want to answer. For instance, in the article in [33], the authors precisely describe behavior since their main goal was to model behavior to follow a vehicle.

If we explore other areas of study, we can find additional definitions of behavior. When referring to a dictionary such as the Michaelis Dictionary, the initial definition for behavior is "the act of behaving" and "the set of reactions observed in an individual in their social environment," which are vague definitions that could be subject to various interpretations. From a transportation science perspective, as described in [5], driver behavior encompasses the set of actions a driver takes to reach their destination.

2.2.2 Representing and Quantifying the behavior

There are several ways to quantify driver behavior, but the representation used must meet certain requirements to be suitable for the behavior prediction model. The goal is to find a representation that can capture the driver's actions in a meaningful way and convert them into quantities that the system can use to predict behavior accurately.

One of the most fundamental ways to represent driver behavior, in terms of immediate sensor data, is to consider the state of the actuators [5]. To this end, various information available in the CAN network can be utilized, such as the accelerator pedal, brake pedal, and engine speed, among others. These quantities can be either real numbers that vary, such as engine RPM, or binary information, such as the status of the brake pedal, indicating whether it is applied or not.

The actuation states produce a series of measurable variables, such as positions, speeds, and accelerations, which can be obtained through sensors and transmitted over the vehicular network. By grouping sets of data containing position, speed, and acceleration information, time sequences can be formed, representing elementary driver actions within a given time interval. For example, a 10-second interval can be analyzed to identify multiple actions performed by the driver during that time. A single acceleration value does not provide a complete picture of the driver's behavior. However, a sudden acceleration throughout the entire interval could indicate a potential aggressive behavior and a tendency towards aggression.

Figure 7 depicts a representation of these moments and the sequence of actuation states during a period when a driver is attempting to pass a vehicle in front of them. The figure illustrates all the different variables and states that occur during this time window.

2.2.3 Driver Behavior in the Literature

Since the 1950s, researchers worldwide have shown a growing interest in understanding and modeling various driver behaviors [34]. Driver behavior analysis provides valuable information that could have many applications in industry and everyday life. For instance, designers of intelligent vehicles need to comprehend driver behavior to design driver assistance systems that can function appropriately in dynamic traffic situations. In contrast, traffic engineers require this information to improve road safety and the reliability of related infrastructure.

The impact of human behavior on specific environments is a crucial factor that influences and contributes to traffic accidents [35]. Treat et al.[35] found that a driver's failure to perceive danger, followed by excessive speed and inattention, are the primary causes of most accidents. Besides human factors, numerous studies have examined the relationship between human characteristics and road safety. Several articles have attempted

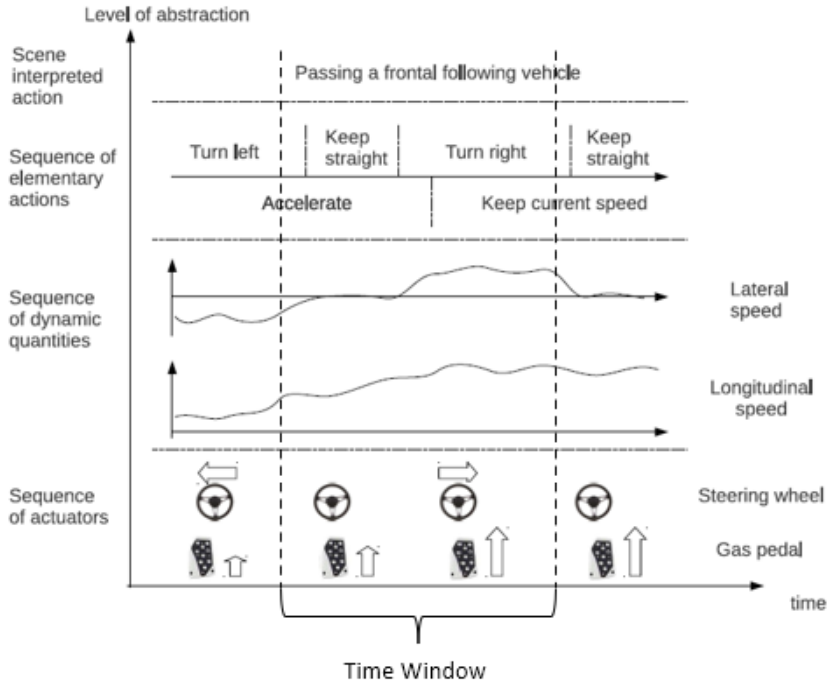

Figure 7. Illustration of possible situation during a trip and time window analysis of the driver action. Adapted from [5]

to develop a model that could define and identify safe and unsafe driving behaviors. According to Miller and Taubman [36], younger drivers are more likely to cause or be involved in traffic accidents than older ones. Kim, Nitz, and Li [37] also discussed that age and driving experience are significant factors in causing accidents. One reason for this is that younger drivers are willing to take more risks or may fail to anticipate dangerous events [[38]]. Due to differences in their ability to anticipate such events, younger drivers may exhibit high-speed behaviors, while their lack of experience in detecting hazardous situations increases the likelihood of causing an accident. Therefore, speed is an essential factor in determining driver behavior.

Events occurring inside and outside the vehicle can also increase the likelihood of an accident. For instance, the use of mobile phones while driving has substantially increased in recent years. Research has shown that distractions inside the vehicle increase the probability and severity of an accident, as the driver's mental capacity needs to alternate between different tasks, resulting in behavioral adaptation such as reducing speed [39]. Once again, speed is a determining factor in characterizing the driver. Other factors that could also influence a trip include passengers, who can greatly affect a driver's concentration and easily distract them [40]. Changing lanes is another example of an external event that may affect a driver's attention.

Drivers exhibit a wide range of driving styles, as there are many factors that come into play when they enter a vehicle and begin their journey. These styles can vary in terms

of how the driver utilizes the accelerator and brake pedals, how frequently the speed of the vehicle changes, and even the speed at which the driver turns the steering wheel. Given this variability, researchers have explored the possibility of using vehicle signals as input to predict driver behavior.

In the research conducted by Aarts and Van Schagen [41], emphasis was placed on the role of vehicle speed in determining driver behavior, considering situations both on highways and in traffic safety. The study discusses the relationship between speed and the probability of accidents, indicating that high vehicle speed not only results in more severe collisions but also increases the risk of accidents.

In Fugiglando's research [8], one of the first analyses of CAN network data was performed with a large sample of drivers in an uncontrolled environment. The study was conducted with 64 different drivers in the city of Ingolstadt, where participants were not given any instructions regarding the route or their driving behavior. The experiment was carried out over a total of 55 days, during which 2418 signals transmitted on the CAN network were collected. The following signals were selected for analysis:

- Brake pedal activation
- Gas pedal position
- Engine speed
- Vehicle speed
- Steering wheel position
- Moment of inertia of steering wheel
- Frontal acceleration
- Lateral acceleration

After analyzing which signals are associated with a direct or indirect intervention between the driver and the vehicle, a signal selection process was conducted. Subsequently, a methodology consisting of four steps, namely feature extraction, normalization, dimension reduction, and unsupervised clustering, was applied to the collected data. The K-means technique was then employed to identify similarities among the various drivers. The results indicate that an optimal number of clusters can be identified using a combination of features, which offer exceptional robustness performance [8].

Constantinescu et al. [42] employed data mining techniques to differentiate and alert drivers about their driving behavior. To gather data, a GPS was developed to collect information from 23 drivers. Statistical methods and a hierarchical clustering algorithm were employed to classify drivers based on their speed, acceleration, and braking behavior.

The researchers also employed the dimension reduction method, Principal Component Analysis (PCA), to reduce the number of variables. The PCA technique enabled them to identify and group the main features. The classification resulted in five groups ranked from non-aggressive to very aggressive.

The research conducted by [43] focuses on analyzing driver behavior at intersections and traffic lights. The authors propose a classification system that categorizes drivers as conservative, normal, or aggressive. The methodology comprises two steps. In the first step, the driver's speed is checked to see if it exceeds the proposed limit by a certain percentage. If the speed is found to be higher, the driver is automatically placed in the aggressive group. The second step involves analyzing the driver's decision to stop or proceed at an intersection. If the driver advances on a yellow or red light, they are also classified as aggressive.

Driver behavior analysis for hybrid vehicles has been explored in the literature [44]. The objective of the authors was to develop a classification system that determines the optimal switching strategy between hybrid energy sources in the vehicle. The drivers were categorized into three groups: conservative, moderate, and aggressive, based on empirical criteria.

The article by [45] proposes a ranking approach to sort drivers based on various sensor input values recorded during their driving. In addition to the data collected while driving on test tracks, the categorization of drivers formed the basis for a cost-effective classification of driving styles.

The research conducted by [46] analyzes drivers based on their fuel consumption levels. The authors classified the drivers into three categories based on a dynamic factor that represents the driver's level of fuel economy. This factor was calculated using the position of the accelerator pedal and the speed of the vehicle. The three categories of drivers were defined as economic, moderate, and sports.

The study conducted by [47] provides evidence that driving style can influence driver safety, particularly if driving style is defined in terms of driver predictability and consistency in performing maneuvers throughout the trip. The researchers utilized sensor data from mobile phones to classify drivers into two distinct categories: aggressive and non-aggressive.

The authors of [48] concur that choosing only two classifications for drivers based on their predictability and consistency in vehicle dynamics is the best approach to maintain generality across all applications.

Based on the research conducted by [49], individuals exhibit varying speed patterns in various traffic scenarios, and these patterns could be utilized to identify driver behavior and other physical and psychological states.

The research conducted by Shi, Zhou, and Qiu [50] analyzed data collected from real-life driving experiences of various drivers. The researchers selected twelve features extracted from vehicle data for the experiment. Initially, the data was reduced using PCA to obtain three main components. Subsequently, a combined algorithm of neural networks and K-means was proposed to classify the three components with a score. The results of the experiment suggested that the combination of these techniques is highly effective in grouping and identifying driver behavior.

In the study conducted by Kalsoom and Halim [51], a comparative analysis was conducted between the K-Means and Hierarchical clustering techniques for the classification of drivers based on vehicle features such as braking frequency, average gear, average and maximum speed, and other related factors. The drivers were grouped into three distinct categories, and the results indicated that the K-means technique outperformed Hierarchical clustering.

The aforementioned researchers classified the driver based on the entire dataset acquired, treating it as a single trip. However, another approach worth exploring involves the use of time windows during the trip to analyze the driver's behavior within each period. This perspective remains underutilized in the literature, and further investigation is needed to fully understand its potential benefits.

In the study conducted by Langari and Jong-Seob Won [52], a method was proposed to classify the driver's style based on the ratio between the standard deviation and the average acceleration, extracted from the acceleration profile within a time window. The driver's behavior was classified as calm if the ratio was less than 50%. If it ranged between 50% to 100%, the driver was considered to have a regular driving style. On the other hand, if the ratio exceeded 100%, the driver was classified as aggressive.

In a study by Murphey et al. [6], time windows were used to determine driver behavior. The study proposed an approach to dynamically classify the driver through jerk profile combined with drivers' road condition statistics in specific time windows. The drivers were classified into three different zones to improve the driver's fuel consumption calculation.

Upon reviewing the existing research, several research gaps and open questions have been identified. Firstly, there is variability in the number of driver behaviors considered across the different studies, with most studies focusing on three distinct behaviors, although this can vary depending on the specific research. Secondly, the selection of parameters to determine driver behavior also varies among the different works. These parameters play a crucial role in characterizing driver behavior accurately. This research aims to address this issue by employing a specific methodology to determine driver behavior, providing consistency and comparability across the analysis. Lastly, the choice of technique utilized for behavior determination varies among the studies; however, it is

noteworthy that K-Means clustering emerges as the most commonly employed technique across the existing research. By examining these gaps and leveraging the strengths of K-Means clustering, this study seeks to contribute to a more comprehensive understanding of driver behavior.

The methodology used in this work to classify the driver will also use dimension reduction and statistical analysis during time windows, taking into account all the references discussed until now. To accommodate the dynamics and different conditions faced by commercial vehicle drivers, four categories for classification will be considered: steady state, calm, normal, and aggressive.

2.3 Clustering

2.3.1 Introduction to Clustering

Machine Learning is a subfield of Artificial Intelligence, commonly referred to as AI, that facilitates the ability of machines to learn by leveraging statistical techniques applied to data. Machine Learning can be categorized into three groups: Supervised Learning, Unsupervised Learning, and Reinforcement Learning.

Clustering analysis is an unsupervised machine learning approach where the primary objective is to group objects in a way that objects within a group are more similar to each other than to objects in other groups. According to [53], clustering is defined as grouping data objects in a dataset based on similarity while maintaining differences from objects in other groups. Clustering is performed for various purposes and applications, but it is primarily utilized for data interpretation and compression [54]. Several clustering applications exist, mainly used for data mining, pattern detection and classification, grouping objects based on similarity or difference criteria, and others [55]. Moreover, clustering has found application in various fields, including image processing [56]. The most commonly used clustering techniques include Model-Based Clustering, Partitioning Clustering, Hierarchical Clustering, and Grid-Based Clustering [57].

2.3.2 Dimension Reduction

In many applications, the excessive number of dimensions can pose a challenge. When utilizing the clustering technique, it is crucial to employ specialized solutions that address the challenges associated with high-dimensional data, as discussed in [58]. Dimension reduction plays a crucial role in various applications, including web dataset search, document classification, and gene expression analysis [59]. Examining and understanding the relationship between different variables becomes challenging when data contains numerous features. This issue not only makes exploratory data analysis difficult but also

affects the performance of machine learning models, potentially leading to overfitting [60]. One approach to address overfitting is to limit the number of dimensions, employ regularization techniques, or increase the amount of training data.

Dimensionality reduction involves obtaining a set of latent variables by decreasing the number of considered random variables. Two primary methods exist in the state-of-the-art: Feature Selection and Feature Extraction [61].

Some advantages of dimension reduction are:

- It reduces the time and storage space required [53].
- Limits overfitting [60]
- Visualization becomes simpler when reduced to low dimensions, such as 2D or 3D.
- Removes multi-collinearity, which helps in the interpretation of model parameters [59]
- Helps avoid the effects of the curse of dimensionality [53].

This thesis aims to reduce the number of dimensions due to the aforementioned advantages.

2.3.3 Curse of Dimensionality

Making a sound decision relies on the available data. As the number of variables or dimensions increases, the amount of required data needed to generalize accurately grows exponentially. To address this issue, several techniques have been developed to overcome the curse of dimensionality. One such technique involves projecting high-dimensional data into a low-dimensional space. A practical example of this reduction is a light projection onto a 3D object, where a 2D shadow appears on the wall. The main benefits of reducing high dimensions include fewer dimensional redundancies and less computational effort. Dimensionality reduction is performed prior to applying clustering techniques.

2.4 Methods for Dimension Reductions

2.4.1 Principal Component Analysis - PCA

Principal Component Analysis (PCA) is a linear feature extraction technique [62]. It is a multivariate statistical method that aims to extract the maximum information from the original dataset variables and represent it in a new set of orthogonal uncorrelated

variables known as principal components [62]. PCA is capable of reducing dimensions in a dataset while performing unsupervised machine learning.

PCA is designed to find a linear combination of variables that summarizes the maximum variance in those variables. The first principal component is the linear combination that explains the maximum data variance. Once the first principal component is determined, PCA searches for the second linear combination that gives the maximum remaining variance, and so on. It is essential to note that the second principal component will be orthogonal to the first component.

In simpler terms, PCA combines input features in a specific way, retaining the most valuable information from all features while discarding the less important ones. It is worth mentioning that the PCA components are linearly independent.

There are three main steps for calculating the PCA:

- Data covariance matrix is computed. To accurately measure the covariance matrix, it is necessary to first normalize the data to have a zero mean and variance units. This ensures that each parameter carries the same weight within the analysis. It is highly recommended to normalize the data before performing PCA; otherwise, variables with high values and variances may dominate the first component, even if they should not.

  In PCA, data variance is calculated from a single random variable, whereas covariance measures how much two random variables are correlated with each other. When two variables are positively correlated, uptrends in one variable correspond to uptrends in the other variable, and when the correlation is negative, they will have opposite trends. The covariance matrix is an array that specifies the covariance between two variables based on the position in the matrix. Its equation 2.1 is given by:

$$\sum = \frac{1}{n-1}((X - \bar{x})^T(X - \bar{x})) \tag{2.1}$$

  Where:

  n: number of data

  x: average of vector values for each X

- Eigenvalues and vectors of the covariance matrix are computed Eigenvectors represent the principal components that determine the direction of the new space obtained through PCA. Eigenvalues, on the other hand, indicate the magnitude of the corresponding eigenvectors. In the proposed application, the eigenvalues will serve as a measure of the variance contributed by each eigenvector.

- The eigenvectors and eigenvalues are utilized to identify the fundamental variable vectors, which are then used to transform the data into a reduced dimension space.

After the eigenvectors and eigenvalues of the covariance matrix are computed, they are used to select the essential variable vectors, and then the data is transformed into these vectors to reduce the dimensions. The eigenvectors are then sorted in descending order based on their respective eigenvalues, with an eigenvector having a corresponding high magnitude in the eigenvalue indicating that the data has a high variance value along that vector in that space. Any feature that can be considered unimportant can be removed if changing its vector does not significantly affect the data. The eigenvector importance list can be used to select only the most critical features and remove the least important ones. The most important features can be selected by analyzing the percentage of explained variance in the variable vectors, which can be up to 80% of the variance of the data in the examples studied. Finally, the data is projected onto the maintained vectors to complete the PCA process.

2.4.2 t-Stochastic Neighbour Embedding - t-SNE

T-Stochastic Neighbor Embedding (t-SNE) is a widely used technique for reducing dimensions and exploring data, first introduced by Van der Maaten and Hinton [63]. Unlike other dimension reduction techniques, t-SNE is capable of handling nonlinear data and preserving local structure, enabling the creation of 2D and 3D data visualizations even for datasets with thousands of dimensions [63].

In the previous section, PCA was introduced as a technique with similar objectives to t-SNE. However, compared to t-SNE, PCA has a limitation in that it only captures linear projections. On the other hand, t-SNE creates a low-dimensional mapping by taking into account the local relationships between data points. This enables t-SNE to capture the nonlinear structures that are present in most datasets. Although other techniques such as Kernel PCA also utilize local structures, t-SNE stands out in practice because it considers both local and global structures [64].

The steps for implementing the t-SNE are as follows:

1. A probability distribution illustrates the correlation among neighboring points in multidimensional space, and a Gaussian distribution is commonly used as a measure of the probability of similarity between these points. One of the essential parameters in t-SNE is perplexity, which indicates the number of neighbors considered for each data point. If the perplexity value is small, the focus is on the local structure, while a larger value emphasizes the global structure. In practice, values between five and fifty are usually effective.

2. In t-SNE, the probability distribution t is used to model similarities between pairs of high-dimensional points, while the probability distribution q is used to model similarities between pairs of low-dimensional points. The technique minimizes the divergence between these two distributions using a gradient descent algorithm. The resulting low-dimensional map preserves the local and global structure of the original data, making it a powerful tool for data visualization and exploration. However, it is important to note that t-SNE is a computationally intensive technique and can take a long time to run on large datasets.

The conditional probability that a point Xi chooses Xj as its neighbor in a multi-dimensional space could be expressed by the formula below, which follows the Gaussian distribution:

$$p_{ij} = \frac{exp(-\left|\left|x_i - x_j\right|\right|^2 / 2\sigma_i^2)}{\sum_{k \neq i}(exp\left|\left|x_i - x_k\right|\right|^2 / 2\sigma_i^2)} \tag{2.2}$$

The conditional probability in low-dimensional space is:

$$q_{ij} = \frac{exp(\left|\left|y_i - y_j\right|\right|^2)}{\sum_{k \neq i} exp(-\left|\left|y_i - y_k\right|\right|^2)} \tag{2.3}$$

The disadvantages of using this technique are:

- t-SNE is a non-deterministic technique, which implies that it is possible to perform multiple simulations of the algorithm and obtain different results in each iteration.
- The different results in each simulation present a problem when there are complex manifolds since it assumes that the local structure of the manifold is linear [65].

2.5 Clustering Methods

The process of clustering involves grouping or selecting data points to identify similar patterns within the data. This falls under the domain of unsupervised learning, and there exist various techniques for performing clustering.

To perform cluster analysis, one must specify a distance metric that quantifies the similarity between two data points [65]. This similarity is also referred to as the dissimilarity measure. The two most commonly used distance measures are the Euclidean and Manhattan distances.

The Euclidean distance is calculated as the distance between two points in two or three-dimensional space using the following equation:

$$d_{euc}(x, y) = \sqrt{\sum_{i=1}^{n}(x_i + y_i)^2} \tag{2.4}$$

As for the Manhattan distance, it is measured along the axles at right angles. The equation below illustrates this distance calculation:

$$d_{man}(x, y) = \sum_{i=1}^{n} |(x_i - y_i)| \tag{2.5}$$

In the following sections, three clustering techniques will be presented, and one will be selected as the primary approach for this research.

2.5.1 K-Means

K-Means is a widely used and popular unsupervised learning algorithm [66] that aims to identify natural groupings or clusters of objects in such a way that objects within the same cluster are similar to each other. In this technique, the parameter k represents the number of clusters that the data should be partitioned into.

The algorithm begins by selecting the number of clusters k. It then initializes the centroids of each cluster, which can be randomly selected or generated from the dataset. The algorithm then proceeds with two main steps:

1. Each cluster is assigned a centroid. The smallest Euclidean distance between the data and the cluster is calculated for each data. Whereas X = x1, x2, ..., xn is the dataset that needs to be grouped into the cluster set S = s1, s2, ..., sk. Data points have d dimensions. In this way, each point x can be represented by the equation below:

$$argmin_{s_i \epsilon S}\left(\sqrt{\sum_{i=0}^{d} s_{i-x_i}}\right) \tag{2.6}$$

2. The next step in the K-Means algorithm is to update the centroids. During this step, the initially selected centroids are recalculated. This is achieved by computing the average of all data points within each cluster. The technique is named after this operation of computing the cluster means. The formula used for centroid recalculation is given by the equation below:

$$c_i = \frac{1}{|S_i|} \sum_{x_i \epsilon S_i} x_i \tag{2.7}$$

The K-Means algorithm performs an iterative process that alternates between two steps until a stopping criterion is met. The criterion can be specified as a maximum number of iterations, minimizing the sum of distances between data points and their assigned cluster centroid, or when no data points change their assigned cluster.

It should be noted that the K-Means algorithm always converges to a solution, as it will stop when one of the stopping criteria is met. However, the solution obtained may not be the global optimum but rather a local optimum. To address this issue, the algorithm can be run multiple times, and the solution with the best objective function value can be chosen. Additionally, statistical analysis can be performed on the various solutions to provide further insights into the clustering results.

Some advantages of K-means are:

- It's easy to implement
- Produces more accurate clusters than other techniques
- It is computationally faster than the hierarchical cluster technique for large data sets and has $O(n^2)$ complexity [67]

Some disadvantages of the algorithm are:

- Scalability sensitive
- The initial estimation of centroids has a direct impact on the result
- It is necessary to define the number of clusters to group the data, which can be difficult for some problems.

2.5.2 Hierarchical Clustering

Hierarchical Clustering is an unsupervised learning technique in which similar data points are grouped in a hierarchical manner [68]. It can be classified into two main types:

- **Agglomerative:** Agglomerative hierarchical clustering is a type of hierarchical clustering in which each data point is initially assigned its own cluster. The nearby clusters are then merged into larger clusters based on a predefined criteria until all data points are assigned to a single cluster. This technique is generally applicable for data sets with few clusters.
- **Divisive:** In this type of hierarchical clustering, all data points are initially considered as belonging to a single cluster and are subsequently divided based on their differences in a top-down approach. Unlike agglomerative clustering, which merges

nearby points to form clusters, divisive clustering recursively splits the dataset until all points are assigned to individual clusters. Divisive clustering is generally used for datasets with a large number of clusters.

Both types are dependent on two main criteria:

- **Distance metric:** measures the separation between two data points, which means that similar points have a smaller distance and different points have a greater distance.
- **Linkage criteria:** determines where the distance is measured between two clusters. There are four main linkage criteria types, defined below for clusters i and j and x being the dataset.
- **Single linkage:** In this criterion, the shortest distance between two clusters is chosen as a reference point in the measurement by selecting two points, which is also known as a nearest neighbor. The distance function can be defined using the following expression:

$$d_{SL}(i,j) = \min_{x_i \epsilon i, x_j \epsilon j} d(x_i, x_j) \tag{2.8}$$

- **Complete linkage:** In this case, the two points that are farthest from each other between two clusters are chosen, making the distance as large as possible. These points are the least similar to each other. The equation for calculating the distance is given by the following expression:

$$d_{CL}(i,j) = \max_{x_i \epsilon i, x_j \epsilon j} d(x_i, x_j) \tag{2.9}$$

- **Average linkage:** in this type, the distance is measured by the average of the clusters, similar to the one used by the k-means technique. The expression below defines this type:

$$d_{AL}(i,j) = \frac{1}{N_i N_j} \sum_{x_i \epsilon i} \sum_{x_j \epsilon j} d(x_i, x_j) \tag{2.10}$$

- **Ward's Linkage:** the objective of this type is to minimize the intra-cluster variance, through the expression below. It is the most common type and it is used for many types of data sets.

$$d_w(i,j) = ||x_i - x_j||^2 \tag{2.11}$$

Hierarchical clustering is known for its ease of implementation, and the resulting structure obtained from applying this algorithm provides more information compared to K-means. Nonetheless, this method comes with some drawbacks. Once the data points are merged, it cannot be reversed, and it is sensitive to good initialization. Furthermore, hierarchical clustering does not handle missing data and requires the specification of the number of groups, which may result in coincidental groups.

2.5.3 DBSCAN

DBSCAN is an unsupervised clustering algorithm that relies on the density of data points to group them together [69]. Its objective is to identify and group data points that are densely packed in close proximity. DBSCAN is known to perform better than centroid and hierarchical clustering techniques in certain scenarios, particularly when the data is not evenly distributed in a linear fashion [65]. For instance, when data points form dense clusters resembling concentric circles, DBSCAN can accurately group them together.

This technique requires the definition of two parameters, namely:

- **Cluster radius:** defines the minimum distance that data will be included within the cluster
- **Minimum number of points:** represents the number of points within the cluster.

When neighboring points are present within the cluster's radius, the cluster expands, and the expansion stops if the criteria are not met, with the points not within the selected set considered as noise. DBSCAN is a robust algorithm, which can be effective with the correct parameters [65].

2.5.4 Choice of Technique

According to the theoretical review conducted, two main approaches can be employed to determine driver behavior: supervised learning and unsupervised learning. Supervised learning is applicable when real-time detection of driver aggressiveness is required, or when the dataset has been trained with labeled instances of calm or aggressive driving behavior. In contrast, unsupervised learning is better suited for the purposes of this thesis, as it involves the extraction and analysis of data, followed by the classification of drivers into distinct behaviors. Among the various unsupervised learning techniques, K-Means stands out as one of the most advantageous, as evidenced by its utilization in other research studies 2.2.3.

In Chapter 2, a comprehensive exploration of different techniques was conducted, ultimately leading to the selection of K-Means due to its distinct advantages in handling

large datasets and effectively clustering data. This technique offers notable benefits, including scalability, ease of interpretation, and flexibility when compared to Hierarchical Clustering and DBSCAN.

One advantage of applying K-Means over Hierarchical Clustering is scalability. K-Means demonstrates excellent scalability, enabling efficient processing of large datasets. This characteristic is particularly valuable when dealing with extensive vehicle data involving multiple drivers. Regardless of the dataset size, K-Means can produce reliable and relatively fast results, making it well-suited for analyzing large volumes of vehicle data. In contrast, Hierarchical Clustering can be computationally expensive and may encounter challenges when dealing with large datasets, leading to longer processing times and potential limitations in terms of scalability.

Another advantage of K-Means is its ease of interpretation. The algorithm is intuitive and straightforward, making it accessible to both researchers and practitioners. Its simplicity facilitates the clustering of extracted vehicle data, allowing for meaningful comparisons and analysis of different driver behaviors. In contrast, Hierarchical Clustering may generate complex dendrograms and require more advanced techniques to determine the optimal number of clusters, making it less straightforward to interpret and apply in practice.

Moreover, K-Means offers flexibility in terms of the number of clusters to be generated. The number of clusters in K-Means is pre-defined, allowing researchers to determine the desired number based on their specific requirements and objectives. This flexibility enables the customization of the clustering process to best suit the characteristics of the dataset and the research objectives. In contrast, Hierarchical Clustering produces a hierarchical structure, making it less flexible in terms of defining a specific number of clusters.

By leveraging unsupervised learning and employing the K-Means algorithm, this research achieves a robust and efficient approach to driver behavior classification based on the extracted data, benefiting from its scalability, ease of interpretation, and flexibility compared to Hierarchical Clustering and DBSCAN.

3 Vehicle Data Extraction

This chapter details all the data extraction performed in the test vehicles. Furthermore, the treatment of the extracted data is also discussed.

3.1 Data Collection

3.1.1 Test Vehicles

As described in Chapter 1, this research was conducted in collaboration with Volkswagen Truck and Bus, a company that develops its products in Brazil. To maintain close relationships with its customers, VWCO lends newly developed vehicles to partner companies for testing in real-world applications. These vehicles are assigned to partners for a short period and are expected to be used in various conditions. Prior to being sent to customers, the vehicles are equipped with a connectivity module called RIO to enable real-time data collection. During the testing period, the engineering department monitors the vehicles' status on a weekly basis through the Customer Engineering area, which compiles and distributes relevant data to other areas of the company.

VWCO offers three vehicle families that are sold worldwide, as shown in Figure 8. From left to right, the models are:

- **Delivery:** This family of vehicles has an application range from 3.5 tons to 17 tons, and they are typically utilized for urban applications, including deliveries within the city.
- **Constellation:** This family of vehicles has an application range from 17 tons to 50 tons and is commonly utilized for various applications, particularly agricultural use, such as sugar cane harvesting.
- **Meteor:** This is an extra-heavy vehicle capable of carrying up to 70 tons and is typically utilized for off-road applications or to travel long distances between logistics centers.

The vehicles that were assigned to be studied in this dissertation were the Delivery and Constellation models, divided as follows:

- **CTV-02:** Delivery model, used in the region of Piracicaba, São Paulo.
- **CTV-04:** Delivery model, used in the capital of São Paulo.

Figure 8. Volkswagen Vehicles: Delivery, Constellation and Meteor

- **VW-421:** Constellation model, used in the rural region of Rio Claro, São Paulo.
- **VW-432:** Constellation model, used in the region of Barueri, São Paulo.
- **VW-433:** Constellation model, used in the Vale do Paraíba region, São Paulo.
- **VW-437:** Constellation model, used in the region of Ribeirão Preto, São Paulo.

The vehicles were loaned to partner companies from November 2020 to February 2021 and were operated on weekdays by their employees. The data acquisition was conducted in an uncontrolled environment, with unfamiliar streets and real traffic situations. Figure 9 provides a visual representation of the origin location for each vehicle in the state of São Paulo, Brazil. The vehicles began their routes from these points and were used for various applications, including the transportation of goods from one distribution center to another.

While the scope of this study primarily centers around commercial vehicles, the findings and methodologies presented can be extended to other vehicle classes, including passenger cars. The utilization of a CAN network for inter-communication and data exchange with various control units is prevalent across most vehicle types. Although specific information may vary among different vehicle models, this research focuses on commonly available data that can be applied to any vehicle. Examples of such shared information include vehicle speed and related parameters. By leveraging this common dataset, the proposed methodologies and insights gained from this study can be adapted and implemented in a broader range of vehicles, facilitating a more comprehensive understanding of driver behavior across the automotive domain.

3.1.2 Data Acquisition System

The Data Acquisition System (DAQ) refers to the process of collecting electrical signals that measure real-world physical conditions, such as those in vehicles, and trans-

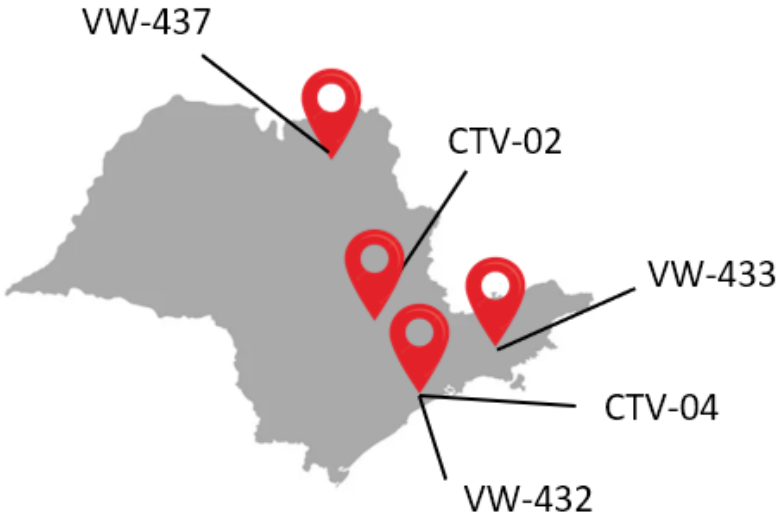

Figure 9. Localization map of test vehicles in the state of São Paulo

mitting them to devices like computers or mobiles for further processing [70]. In this study, the vehicles were equipped with a connectivity module called Rio to acquire this information for engineering purposes. This module comes with two primary communication interfaces: CAN and cellular telephony, which uses a SIM card. The assembly of Rio inside the vehicles before sending them to customers is shown in Figure 10. The module is assembled inside the cabin and has an LED that indicates the box's status.

Figure 10. Assembly of Rio electronic module on VWCO plant

The Rio module is connected to the Powertrain CAN, which serves as a communication interface between the onboard computer and the engine control module. Most of the data inside the vehicle, such as vehicle speed, engine speed, and fuel consumption,

among others, are transmitted through the CAN interface studied in section 2.1.1. Once the vehicle is turned on, the module is activated and initiates the data collection process. At pre-defined intervals, the module sends the information to the nearest cell phone antenna. The data is then transmitted to VWCO's internal server and stored in the Amazon AWS cloud, which is accessible through a program used exclusively within the company.

Drivers are aware of the ongoing data collection process, wherein fleet owners require a formal agreement to extract and utilize truck data for additional applications. Throughout the journey, the RIO Box, responsible for gathering relevant information, illuminates a green indicator light to signify data extraction. Nevertheless, based on the conducted interviews, it has been observed that drivers predominantly prioritize road safety and reaching their destinations promptly, often inadvertently neglecting their awareness of continuous monitoring.

Data were extracted directly from the server and automatically compiled into an Excel spreadsheet generated by the program at the end of each day. The spreadsheet includes a column for each feature extracted from the vehicle, with the last column containing raw data taken from the CAN network.

The information available in each of the worksheets was:

- Date and time
- Total odometer
- Total engine hours
- Total fuel used
- Instantaneous fuel consumption
- Vehicle speed
- Engine Speed
- Oil Pressure
- Engine charge pressure
- Brake pedal
- Parking brake
- Clutch pedal
- Gas pedal position
- Engine retarder

- Current gear
- Average fuel consumption
- Coolant temperature
- Engine torque in percentage

When collecting data, it is essential to consider both external and internal factors that may affect the results. Therefore, to ensure a more accurate analysis, the data must undergo appropriate treatment. This treatment also applies to data extracted by electronic modules inside the vehicle.

3.2 Data Treatment

In order to apply data treatment, it was used Python programming language and the Jupyter Notebook tool to conduct this phase. Jupyter Notebook is a free software with a straightforward visualization that is easy to use, and its installation on the Anaconda platform enables its use on a local machine utilizing an available browser.

The data processing was performed following the steps outlined below:

1. **Step 1:** The initial step involved excluding the first line from the data files, which contained the table title and the date of data collection.
2. **Step 2:** Subsequently, with the first line removed, the data tables were concatenated into a single table for each vehicle. The process utilized the glob() function, the Openpyxl library, and functions from the Pandas library.
3. **Step 3:** The third step involved the removal of units within the data cells to eliminate any influence on the following calculations. This was accomplished using the removeunit() function.
4. **Step 4:** As a subsequent step, certain data was converted from object type to float64 using the astype() function to enable joint analysis with other data.
5. **Step 5:** Finally, the lines with missing data were eliminated using the dropnas() function. It was observed that approximately 40% of the vehicle data was missing due to the fact that not all features are transmitted on the CAN network simultaneously or in the same message cycle. As the focus of this section was on analyzing the various features, only the rows containing complete data for all features were selected.

After performing the aforementioned five steps in all vehicles, individual data files were obtained for each vehicle listed above. These files encompass data from diverse days, trips, and drivers, and contain approximately 400,000 lines of data each. The resultant files pave the way for the upcoming feature selection phase, which will be explained in Chapter 4.

4 Data Features Selection

4.1 Features and driving parameters

Analyzing and determining driver behavior is a complex task that takes into account several external and internal factors. As previously detailed in the last chapter, 18 features were extracted from the vehicle's CAN network. This chapter aims to discuss the main features extracted from the CAN network and their general impact on driving behavior.

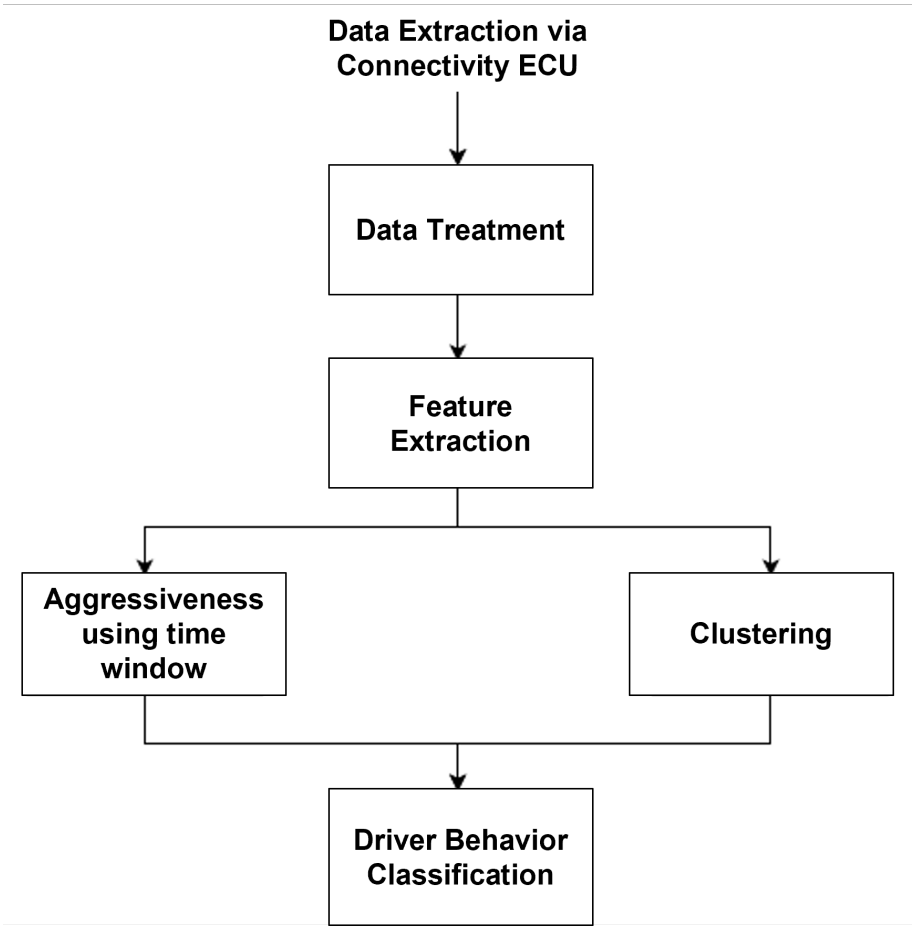

Figure 11. Brief overview of this thesis methodology for driver behavior and classification

The methodology to be applied in this chapter and the following is summarized in Figure 11. The process for the final result of this work occurs in two steps. In the first step, relevant information collected on the vehicle's CAN network is analyzed and extracted since many are complementary. Extracting features from data collected in continuous time is considered challenging since some features are unavailable at all times in the time series and require extra processing for the information to be relevant [65]. The second

step involves finding the driver's aggressiveness in time windows and grouping the data according to predefined clusters. For example, the gear indication is a feature where the current gear status is sent at each time interval of the message cycle. The current gear not only provides relevant information to determine the driver's aggressiveness, but it can also be used to analyze the comparative data with the last moment and to verify if the gear change was made at the right moment and duration. Although it would be possible to use supervised learning to determine the driver's aggressiveness, the complexity mentioned in the theoretical review of determining a profile and exact rules for the driver's behavior makes this proposal of the time windows technique and unsupervised learning a more suitable approach to cluster the data sets.

4.1.1 Software Environment

From this point of work, R Studio will be used as a basis for development. The version and libraries used are shown in table 1.

Table 1. Comparative table of extracted features and their relevance for driver behavior

Name	Description	Version
R Studio	R Studio is an integrated development environment (IDE) for R.	2021.09.02
tidyverse	Library for better data visualization	1.3.0
readr	Library used for reading flat files, like csv	-
ggbiplot	Library used for plot and data visualization	-
corrplot	Library used for correlation plot	-
PCA tools	Library used for principal component analysis implementation	-
dplyr	Library used for data manipulation	-
data.table	Library used for handling large groups of data	-

4.1.2 Preliminary Analysis of Features

The initial step in analyzing the features involved determining if they were driver-dependent or simply vehicle-related responses to driver actions. For a feature to be considered driver-dependent, it must have a controller that directly controls the result read on the CAN network. Conversely, vehicle-dependent features are not directly controlled by the driver and are instead a response to some existing actuation on a controller. Driver can influence in this information, however it would take a long time to see an influence of his behavior in these parameters. This analysis was necessary to understand which information in the network influences driver behavior. Additionally, a relevance column was created to categorize the information's significance in calculating driver aggressiveness as low, medium, or high. This led to the creation of a comparative table 2.

The initial aspect that requires analysis and refinement is the date and time data. The network data specifies the time when the information was obtained. In order to

Table 2. Comparative table of extracted features and their relevance for driver behavior

Feature	**Dependence**	**Relevance**
Date and hour	Vehicle	medium
Total odometer	Vehicle	medium
Total engine hours	Vehicle	low
Total fuel used	Vehicle	low
Instantaneous fuel consumption	Vehicle	medium
Vehicle speed	Vehicle	high
Engine speed	Vehicle	medium
Oil pressure	Vehicle	low
Engine charge pressure	Vehicle	low
Brake pedal	Driver	high
Parking brake	Driver	low
Clutch pedal	Driver	medium
Gas pedal position	Driver	high
Engine retarder	Vehicle	low
Current gear	Driver	low
Average fuel consumption	Vehicle	low
Coolant temperature	Vehicle	low
Engine torque	Vehicle	medium

conduct a comparative analysis within time windows, it is essential to establish a reference point of time zero, originating from the start of data collection. Subsequently, the time difference is calculated with respect to the previous value. This approach facilitates the linearity of time and provides significant information for comparative analysis. The following equation illustrates the computation of the time difference:

$$timedif(i) = ((time_{i+1} - time_i) * 86400) + timedif(i-1) \tag{4.1}$$

After the previous step, the subsequent analysis should be performed using the data containing the eighteen relevant features extracted for this study. To conduct this analysis, a selection was made from the previous table of variables that could most significantly influence driver behavior, specifically those with medium or high relevance. Following this preliminary selection, a correlation analysis was carried out between the selected features using the "corrplot" R library.

The Figure 12 and 13 illustrate the found results.

After conducting the first correlation analysis, it can be inferred that there exist three primary groups of information:

1. **First Group:** The first group consists of time and total odometer. These variables are directly proportional to the driver's driving time and will serve to guide and

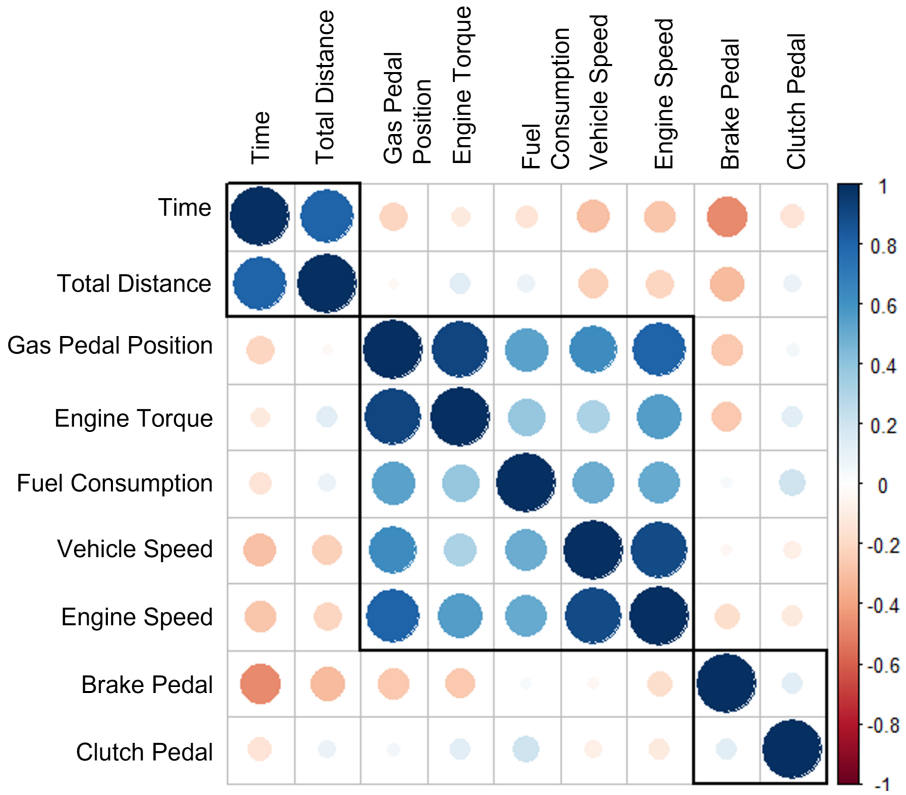

Figure 12. Correlation plot of selected features

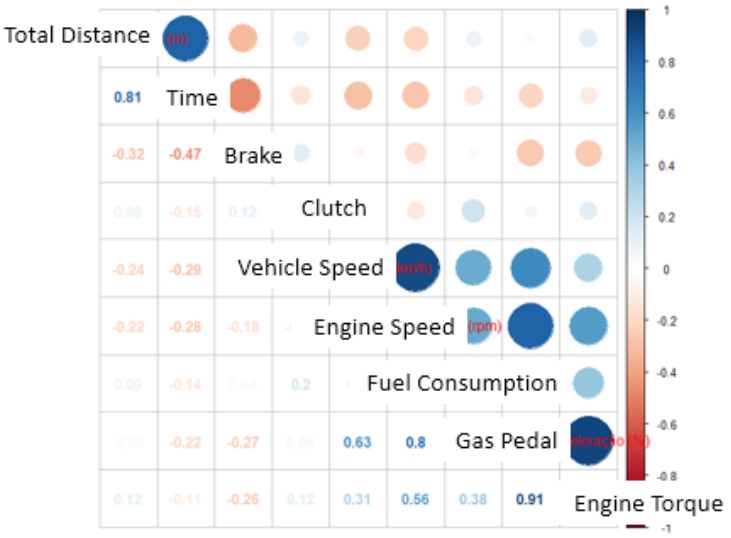

Figure 13. Correlation plot with correlation values of selected features

apply the time windows in the calculation.

2. **Second Group:** The second group includes gas pedal position, engine torque, fuel consumption, vehicle speed, and engine speed. All these variables are positively correlated with each other, as manipulating one of these variables will have a direct impact on the others.

3. **Third Group:** The third group comprises brake and clutch pedal data. This information is primarily related to the driver's braking moments, which may be sudden or gradual.

In order to proceed with the analysis of driver aggressiveness using time windows, three sources of information will be taken into consideration, one from each group. The features that were selected after applying the second filter are time, vehicle speed, and brake pedal. Section 4.1.3 will discuss each of these three sources of information in detail and analyze the insights that can be obtained from them.

4.1.3 Time

The concept of time is fundamental in physics, and it denotes the duration between two events or moments. In the context of data analysis, time can refer to the sequence of events or the time of occurrence of specific events or data points.

In data analysis, time is crucial because it enables the study of trends and patterns over time. Time-series data analysis involves analyzing data collected over time to identify trends, cycles, and other patterns. By analyzing how a variable changes over time, data analysts can make predictions about future behavior or estimate the impact of a particular event on future outcomes.

Moreover, time is significant in data analysis because it allows the use of advanced analytical techniques, such as time-series forecasting, which predicts future values of a variable based on historical data. Time-series forecasting finds applications in finance, economics, engineering, among others, to make predictions about future trends and inform decision-making.

Additionally, time is often incorporated as a factor in machine learning models, such as decision trees and random forests, to predict future outcomes based on past events. Incorporating time as a factor in these models can improve the accuracy of predictions, enabling data analysts to make more informed decisions.

In this dissertation, time is measured in seconds and collected right after the moment when the ignition of the vehicle is on. The time difference between two data samples was used to calculate time. The first collected sample was defined as time 0 - the beginning of the trip, where all data starts to be collected.

Time is crucial to analyze the windows and compare driver actions inside these windows. Using time as a variable, it will be possible to analyze driver behavior as a "movie" and not only as a single "picture".

4.1.4 Vehicle Speed

Vehicle speed is a crucial parameter that helps in determining a driver's profile. It is defined as the distance traveled per unit of time, and it is measured by an internal speedometer in the vehicle in units such as kilometers per hour (km/h) or meters per second (m/s). This information can be directly extracted from the data analyzed in the previous chapter.

Studies have shown that accident risks increase at higher speeds [71]. Moreover, increasing legal speed limits further enhance the risk of fatalities [72]. However, when analyzing the speed variable, it is essential to consider not only the number itself but also the density of speed over time. A driver who maintains a high average speed for an extended period may not be classified as aggressive. Therefore, the distribution of speed density over time can be utilized to examine the smoothness in driving quality.

Figure 14 illustrates the speed data of two different drivers. In this example, the first driver had greater variations in speed over time, which could be explained by frequent acceleration and deceleration (perhaps due to traffic). The second driver, on the other hand, maintained a more constant speed over a longer period.

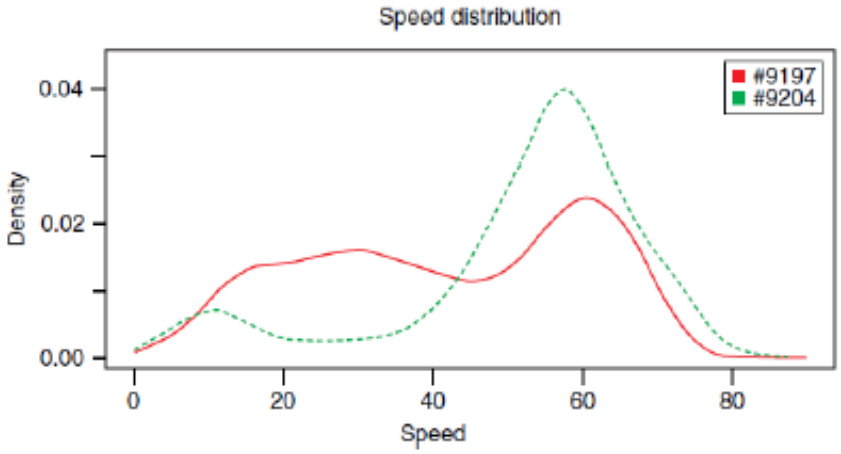

Figure 14. Vehicle Speed Distribution of two different drivers. Extracted from [20-50]

Furthermore, applying statistical techniques to the speed signal can provide additional insights, such as measures of central tendency (mean), variability (standard deviation), and range (minimum and maximum values), as well as measures of shape such as kurtosis and skewness.

Skewness is a statistical measure that indicates the symmetry of a dataset. A perfectly symmetrical dataset will have a skewness value of 0 when plotted on a histogram.

Kurtosis, on the other hand, is a measure of the peakness or flatness of a dataset. There are three classifications of kurtosis: platykurtic, mesokurtic, and leptokurtic. In the case of the speed dataset, kurtosis can be used to classify a driver's style as either urban or highway. A driver with a kurtosis greater than 3 is considered to have a highway style, while a driver with a kurtosis less than 3 is considered to have an urban style [73].

4.1.4.1 Acceleration and Deceleration

An aggressive driving style is often characterized by sudden changes in acceleration, both positive and negative. The unit of acceleration is typically measured in meters per second squared (m/s^2). Sudden changes in acceleration can be obtained by differentiating the speed signal, as shown in the equation 4.2 below:

$$a = \frac{dv}{dt} \tag{4.2}$$

The acceleration of a vehicle has a considerable impact on the driver's conduct, particularly in terms of driving performance, safety, and fuel efficiency.

Regarding driving performance, acceleration influences the driver's capability to overtake other vehicles, merge onto highways, and respond to changing traffic conditions. Drivers who can accelerate smoothly and efficiently can enhance their driving performance, decrease stress, and feel more in command of their vehicle.

Concerning safety, acceleration also plays a crucial role in preventing accidents. Drivers who can accelerate quickly and smoothly can avoid collisions by safely merging into traffic, passing other vehicles promptly, or maneuvering around road obstacles. However, rapid acceleration can increase the risk of accidents, particularly in situations where the road is wet, icy, or otherwise slippery.

Moreover, acceleration can affect fuel efficiency. Drivers who accelerate too quickly or frequently can consume more fuel than necessary, reducing their vehicle's overall fuel efficiency and increasing their operational costs. Conversely, drivers who accelerate gradually and smoothly can enhance their fuel efficiency and reduce their carbon footprint.

For commercial vehicles, an acceleration value exceeding 4 m/s^2 is considered aggressive. This value was established as the driver's aggressiveness threshold for this category of vehicles by CONTRAN 667/17.

4.1.4.2 Jerk

Jerk refers to the rate of change in acceleration, which can be calculated by taking the time derivative of acceleration or the second derivative of velocity. The unit of jerk is m/s^3. This feature is important as it can indicate how sudden the changes in acceleration are. A study referenced in [6] suggests that jerk is a more effective feature than acceleration when it comes to driver classification. The equation 4.3 describes the formula to calculate jerk.

$$j = \frac{da}{dt} \tag{4.3}$$

Jerk is a significant factor that can influence driver behavior, particularly in terms of driving comfort and safety. The rate of change in acceleration can cause discomfort to passengers and increase the risk of accidents. Jerk is measured in units of m/s^3 and can be calculated by taking the time derivative of acceleration or the second derivative of velocity.

When a vehicle accelerates or decelerates rapidly, the sudden change in speed can cause discomfort for passengers. For instance, if a vehicle accelerates too quickly from a stop, passengers may feel thrown back into their seats, causing discomfort and even pain. Similarly, if a vehicle brakes too hard, passengers may lurch forward, leading to discomfort and even injury.

High jerk values can also affect a driver's ability to control the vehicle. Rapid changes in acceleration can make the vehicle unstable and challenging to control, especially in adverse weather conditions or on slippery roads.

Moreover, high jerk values can cause driver fatigue and reduce reaction times, increasing the risk of accidents. The constant jolting and sudden changes in acceleration can be mentally and physically tiring, resulting in reduced concentration and slower reactions to changing road conditions.

The threshold values for jerk that indicate aggressive driving behavior are discussed in Chapter 5.

4.1.5 Brake Pedal Position

The brake pedal position, also referred to as the brake light switch, is a vital element of a vehicle's braking system. Positioned beneath the brake pedal, it is responsible for activating the brake lights when the pedal is depressed.

Once the brake pedal is engaged, the brake pedal switch is activated, and an electrical signal is sent to the brake lights, causing them to illuminate and indicating to

other drivers that the vehicle is slowing down or stopping. This function contributes to enhancing safety on the road and mitigating the risk of rear-end collisions.

In addition to its role in activating the brake lights, the brake pedal switch may also serve as a means of controlling other vehicle systems, including the cruise control, transmission shift interlock, and anti-lock braking system (ABS). Moreover, it can provide crucial data to the vehicle's onboard computer, which can aid in enhancing fuel efficiency and minimizing emissions. In the present dissertation, such information is transmitted to the cloud and can indicate whether the driver's behavior was aggressive or not.

In summary, the brake pedal switch is a relatively small yet fundamental element of a vehicle's braking system, which significantly contributes to improving road safety.

4.1.6 Comments about Selected Parameters

The selection of three key dimensions, with a particular emphasis on speed and its related features, was made to comprehensively analyze and characterize driver behavior, as these variables exhibit the highest variance in the collected data. For instance, engine speed, which is closely correlated with vehicle speed, was intentionally omitted from the analysis to avoid redundancy and prevent it from unduly influencing the results. It is worth noting that while there are other parameters within the vehicle that could potentially be extracted, the decision was made to focus on common information shared across all vehicle types. By prioritizing these common parameters, the findings and methodologies presented in this study can be readily applied to a wide range of vehicles due to their simplicity and ubiquity.

However, it is important to acknowledge certain limitations. Firstly, the year and model of the vehicles involved in the study must be taken into account. Considering the Brazilian legislation, the implementation of stability control systems is mandated for forthcoming years in commercial vehicles. This introduces an important additional sensor to the network—the steering angle sensor—which offers the potential to measure driver behavior related to steering input and detect signs of driver fatigue or drowsiness during traffic. It should be noted that this particular sensor was not considered in the present research due to its absence in the studied vehicles. The incorporation of such sensors in future studies holds promise for enhancing the analysis of driver behavior and expanding the scope of investigations.

5 Time Window Scoring and Clustering of Driver Behavior

In this chapter, the final step of this master thesis will be discussed: the classification of the driver in aggressive behaviors and grouping driver patterns through unsupervised learning.

5.1 Time Window Scoring

In this section, an algorithm based on time windows is presented to assign an aggressiveness score to the driver. The first step to be taken is to determine the number of categories the algorithm uses. As established in Chapter 2, four categories will be utilized, namely:

- **Safe Driving:** This driving behavior can be described as safe, where the driver takes measures to anticipate the movements of other drivers in traffic. They abide by speed limits, and avoid sudden acceleration or braking while driving.
- **Normal Driving:** This driver profile is characterized as moderate driving, with controlled acceleration and braking.
- **Unsafe Driving:** This driving behavior is highly aggressive and poses a danger to others on the road. The driver tends to make sudden acceleration and braking maneuvers, disregard speed limits and fail to anticipate potential hazards.
- **Steady State:** When the vehicle is stationary, i.e. speed is zero, the driver's behavior will be considered steady state as it does not impact traffic.

It is worth noting that a driver's driving style may not be consistent throughout the entire trip. It is common to observe all three types of driving behaviors in the same trip due to the many factors that affect driving. Therefore, the main objective of this algorithm is to identify changes in a driver's behavior and determine if aggressive behavior was predominant during the trip.

Various characteristics can be used to identify aggressive driving behavior on the road, such as excessive speed, sudden acceleration, and unnecessary braking. To measure these characteristics, jerk will be considered, which represents the variation in acceleration, as discussed in the previous chapter. This feature is utilized because acceleration indicates how a driver increases or decreases speed, whereas jerk indicates how the vehicle

accelerates or decelerates. The Figure 15 illustrates the variation of velocity, acceleration, and jerk over time, highlighting how jerk is more stable than acceleration and measures points where acceleration has a significant variation.

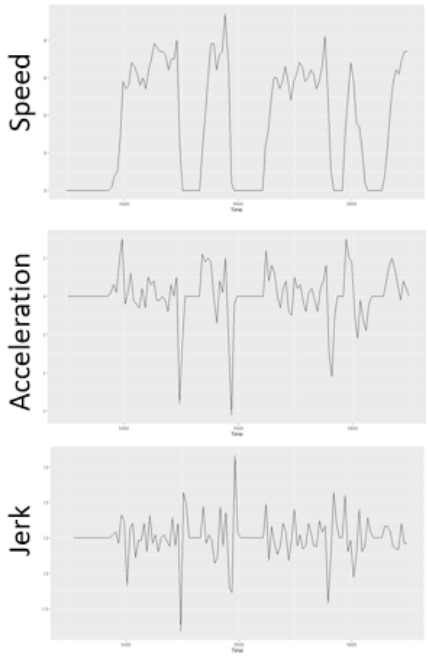

Figure 15. Comparison between speed, acceleration and jerk in a given period

The methodology used for jerk comparison was extracted from the work of Murphey et al [6]. In this article, the equation 5.1 was proposed to determine whether the driver fits as aggressive or not:

$$\gamma = \frac{SD_J}{\bar{J}} \tag{5.1}$$

The value obtained above signifies the jerk feature and is calculated by taking the ratio between the standard deviation of the jerk and the average jerk value for the road. According to the authors, this value needs to be updated based on the road that the driver travels on. In this thesis, the data were collected in uncontrolled environments, and there is currently no way to determine the road condition using the extracted data. Therefore, an average jerk value from all types of roads will be considered, which has been calculated to be 0.2732 based on the mean value of the calculated jerk in [6]. Table 3 demonstrates the found values of jerk for each road type in Murphey et al.'s research [6]. The last row in the table displays the average value for all types of roads.

To perform the calculation, the following steps should be followed:

Firstly, the jerk value is calculated for each time window. The time windows are a parameter in the program, and an analysis will be carried out later on the optimal window size for the data.

Table 3. Average Jerk for each road. Adapted from [6]

Drive Cycle	**Average Jerk (m/s³)**
Freeway A	0.2131
Freeway B	0.2126
Freeway C	0.2258
Freeway D	0.2075
Freeway E	0.2401
Freeway F	0.3096
Ramps	0.2925
Art-AB	0.2580
Art-CD	0.2825
Art-EF	0.2460
Local	0.2439
Mean	0.2732

The algorithm proposed in this section analyzes a driver's behavior using a time window technique, which involves comparing the driver's actions during a time interval (t). At any given time window (t), the classification algorithm extracts the jerk variation during the proposed time and divides it by the average jerk of a highway mentioned above. The resulting ratio is used to classify the driver into one of the three predefined categories in this dissertation. The driver's next classification takes place at the next time delta t, which begins shortly after the end of the previous time window. At that moment, the algorithm recalculates the variation of the acceleration in the new window to classify the driver in real-time. An important parameter in the analysis of driver behavior is the window size, which needs to be optimized. Therefore, several parameter value variations will be compared to determine their impact on the general behavior of the driver.

The step-by-step of the developed algorithm is proposed as follows:

1. To begin the calculation, select the appropriate time window size, taking into account the duration of each behavior exhibited by the driver. Four different window sizes will be compared: 10, 15, 20, and 25 seconds. These window sizes were arbitrarily chosen based on the author's professional experience with trucks.

2. Next, verify that the vehicle speed is zero, indicating that the vehicle is stationary. If the vehicle is stationary, there is no need to calculate acceleration and jerk.

3. Calculate the acceleration and jerk during the predetermined time period.

4. Determine the gamma value, which represents the ratio between the driver's jerk and the average value accepted by a standard driver.

5. Perform a comparative analysis between the gamma value and the defined aggressiveness parameters to classify the driver's behavior.

For the final step, it is necessary to establish the comparison values that will be used to determine whether a driver has sudden acceleration in their vehicle. To accomplish this, a government-defined standard for what constitutes sudden acceleration/deceleration will serve as a basis.

In 2017, the National Traffic Council (CONTRAN) in Brazil, the regulatory body responsible for managing traffic regulations in Brazilian territory, published legislation mandating the implementation of emergency braking signals. To do so, they needed to define what constituted an emergency braking event.

According to the resolution, acceleration values were established based on vehicle class. The system in question should activate a signal for the vehicle behind to indicate that a braking event above the allowed limit has occurred. For passenger vehicles, sudden acceleration above 6 m/s^2 is considered. For commercial vehicles, due to their size and impact on traffic, acceleration above 4 m/s^2 is already considered an aggressive maneuver by the driver.

The value of 4 m/s^2 will be utilized to identify aggressive driving behavior. Acceleration between 2 m/s^2 and 4 m/s^2 will be considered a normal driving style, while below 2 m/s^2 will be categorized as calm driving.

Thus, the comparison intervals and their corresponding acceleration values have been defined. However, the determination of an aggressive jerk value for the driver is still pending. According to Murphey's study [6], an acceptable level of aggressiveness would be above 1.0 and between 0.5 and 1.0 for normal driving style. Nevertheless, this research was conducted on passenger vehicles and cannot be directly applied to commercial vehicles. Hence, the same proportion of 1.5 between sudden acceleration of a car and a truck, as defined by CONTRAN, will be used for commercial vehicles. Therefore, the following values will be considered for jerk:

- Gamma $>$ 0.67, will be an aggressive driver
- 0.33 $>$ Gamma $>=$ 0.67, will be a normal driver
- Gamma $<=$ 0.33, calm driver

Having established the initial parameters, assuming that the aggressiveness calibration will be maintained henceforth, the first simulation of the algorithm can be conducted on a dataset extracted from Driver 1. Upon completion of each simulation, a graph will be generated that illustrates the driver's speed throughout the total time period, with vertical lines indicating instances where the driver's driving style changed. Red vertical lines will denote instances where the driver was driving aggressively, blue lines will signify instances of proper driving behavior, and green lines will indicate periods where the driver can be considered to be driving calmly in traffic.

Moreover, moments when the vehicle was stationary were also identified, during which no driving behavior could be inferred. These moments will be represented by yellow lines. In order to facilitate visualization and enable identification of transitions in the driver's behavior, lines will only be represented when the algorithm detects a change in driving style. Therefore, it must be considered that moments that are not represented by any vertical line indicate that the driver is maintaining the driving style represented by the previous line.

Various time windows were simulated to determine an optimal value for the analysis of other drivers. Initially, time windows of 10, 15, 20, and 25 seconds were evaluated, and the algorithm was recompiled accordingly. Results from these simulations are presented in Figures 16, 17, 18, and 19.

As an example, Figure 18 shows the results obtained from a 20-second time window analysis. A driving time of 5300 seconds was extracted from the beginning of the trip until the 5900-second mark, which corresponds to 10 minutes of driving. During this time, the driver exhibited 16 transitions in driving style, ranging from aggressive to calm. The analysis revealed that the driver drove in a city or a region with traffic, as the vehicle's speed was frequently interrupted. An instance of aggressive behavior occurred near the 5400-second mark, where the vehicle's speed increased from 0 to 30 km/h within a time window of 20 seconds. The algorithm detected an acceleration variation greater than 0.67, identifying this as an aggressive profile. Overall, the analysis identified 9 aggressive driving windows, 12 normal driving windows, 2 calm driving windows, and 6 windows where the vehicle was stationary. Consequently, the driver displayed aggressive behavior during approximately 31% of the driving time.

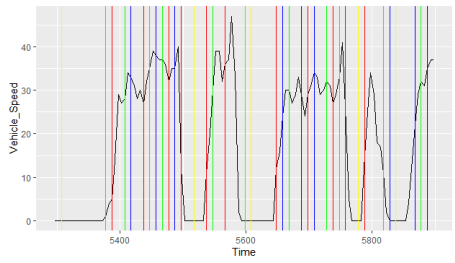

Figure 16. Driver classification in time window of 10 seconds

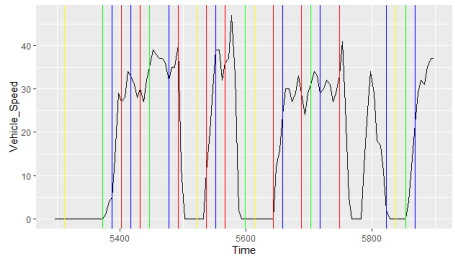

Figure 17. Driver classification in time window of 15 seconds

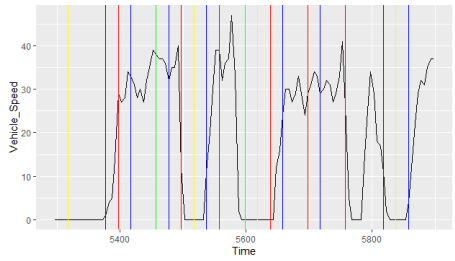

Figure 18. Driver classification in time window of 20 seconds

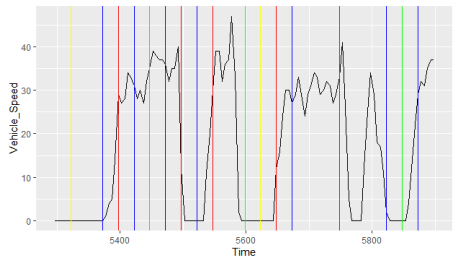

Figure 19. Driver classification in time window of 25 seconds

5.2 Clustering using K-Means

In Section 5.1, a method was discussed for assigning a score to the driver and determining the periods of the trip during which they were aggressive or not. This section introduces another technique for grouping the extracted data into patterns of behavior using unsupervised machine learning, specifically the K-Means algorithm. No initial assumptions were made about trips, drivers, or the extracted data.

When applying unsupervised techniques, it is common to perform visualizations of grouped data in 2D or 3D space. However, since a total of 18 vehicle features were extracted, it is not possible to recreate a space with all of them included for grouping. Therefore, a technique for dimension reduction called Principal Component Analysis (PCA) is necessary before performing the grouping. However, before PCA can be applied, it is crucial to normalize the data due to the significant difference in magnitude between the features. For instance, vehicle speed can reach values up to 100 km/h, while the use of the brake pedal ranges from 0 to 1. The standard normalization method, described in Equation 5.2, was used, where μ represents the mean and σ represents the standard deviation of x.

$$z = \frac{x - \mu}{\sigma} \tag{5.2}$$

The Figure 20 illustrates the steps to perform clustering:

Data normalization is done using the scale command in R Studio.

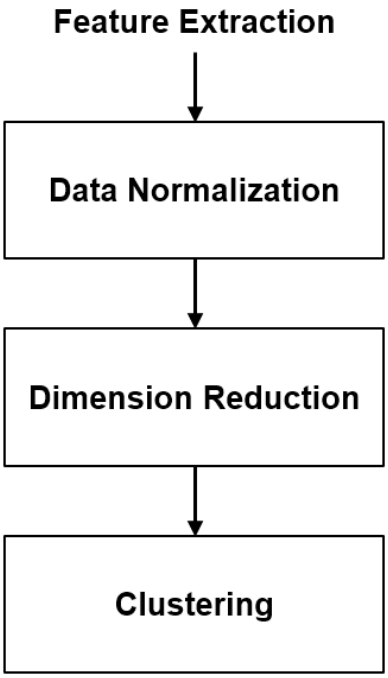

Figure 20. Necessary Steps for Clustering Method

5.2.1 PCA applied in extracted data

By applying PCA to the extracted data, the number of features is reduced for some major components. As explained in section 2.4.1, the technique consists of assigning a component to each dimension and identifying how many percent of variance each component corresponds to.

The technique was applied in R studio software, with the support of the library factoextra and ClusterR. After applying the PCA to a vehicle dataset, the graphs below can be obtained. It is important to note that with only two dimensions it is possible to obtain almost 60% of the variance of all the data. In other words, this means that the 9 dimensions can be represented by the first two principal components.

The summary of the main components, which account for over 60% of the variation in the data, is presented in Table 4. The highest coefficients for PC1 are gas pedal position, engine torque, engine speed, and vehicle speed. On the other hand, for PC2, the highest coefficients are time and odometer. The position of the components in the two-dimensional projection is determined by values with greater magnitude. It is important to note that the set of features grouped in the components is the same as that identified in the correlation plot, which further confirms the accuracy of the data.

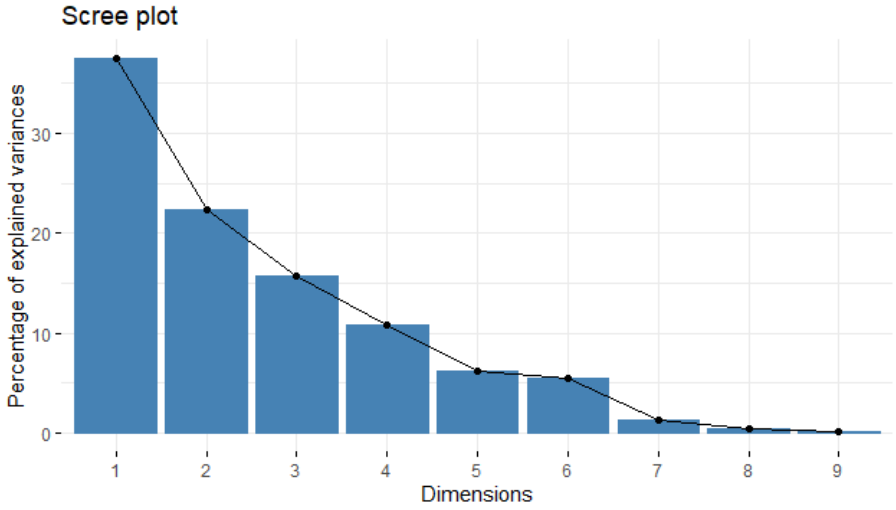

Figure 21. Variance of dimensions for vehicle dataset

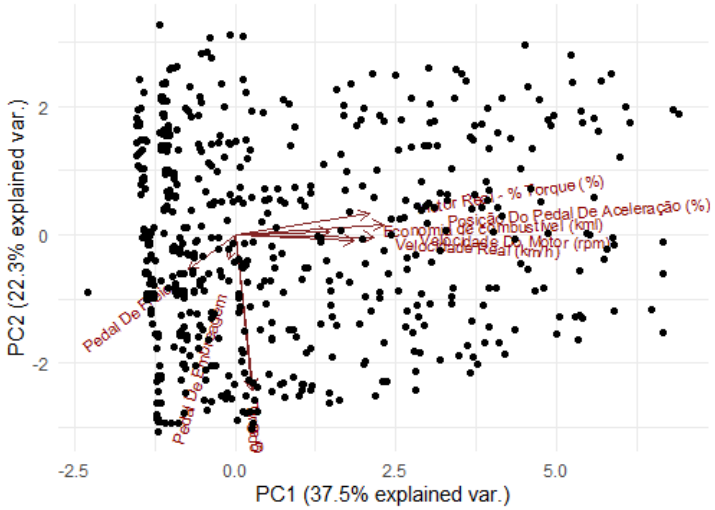

Figure 22. Principal Component Analysis - components visualization with data

Table 4. Summary of principal components x features

Feature	PC1	PC2
Time difference	0.06118189	-0.68898228
Total odometer	0.05746766	-0.69084891
Instantaneous fuel consumption	0.32197797	0.01464038
Vehicle speed	0.40863637	-0.02621686
Engine speed	0.47415214	-0.01071482
Brake pedal	-0.16098384	-0.15584205
Clutch pedal	-0.02362373	-0.11068030
Gas pedal position	0.50861271	0.04492985
Engine torque	0.46078166	0.09197375

5.2.2 Applying K-Means for clustering

After applying the PCA technique and reducing the dimensionality of the data, the K-Means unsupervised learning technique can be applied. However, before applying the K-Means algorithm, it is necessary to tune the method's parameters, namely the number of clusters, k. Choosing the right value for k is crucial for obtaining a good result from the algorithm. If the number of clusters is too small, some groups may be merged, and relevant insights may be lost. Conversely, if k is too large, the data will be overly fragmented, and it will be difficult to gain meaningful insights.

To determine the optimal value of k, the elbow method is commonly used. This method involves plotting a graph to identify the elbow point, which represents the optimal number of clusters. Figure 23 displays the result obtained after applying this method.

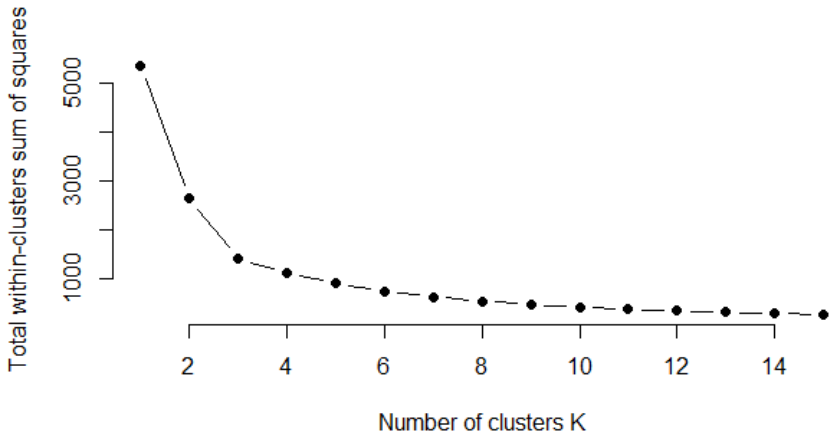

Figure 23. Elbow Method - Finding optimal number of k clusters

The elbow point in the graph indicates that the optimal number of clusters is 3. Therefore, the parameter k will be set to 3 for the K-Means algorithm. By applying the K-Means algorithm with the equations established in Chapter 2, the resulting figure 24 shows that the clusters found are well defined and assume a circular shape. The red cluster, cluster 1, represents moments when the driver was driving normally with little use of the accelerator pedal. The green cluster, cluster 2, represents times when the vehicle was stopped for a delivery. Finally, the blue cluster, cluster 3, represents moments when the driver was accelerating and driving at high speed. Although not all moments in this cluster can be classified as aggressive driving, it is possible to say that moments of aggression occur in this region. The size and averages of each cluster are shown in the table below:

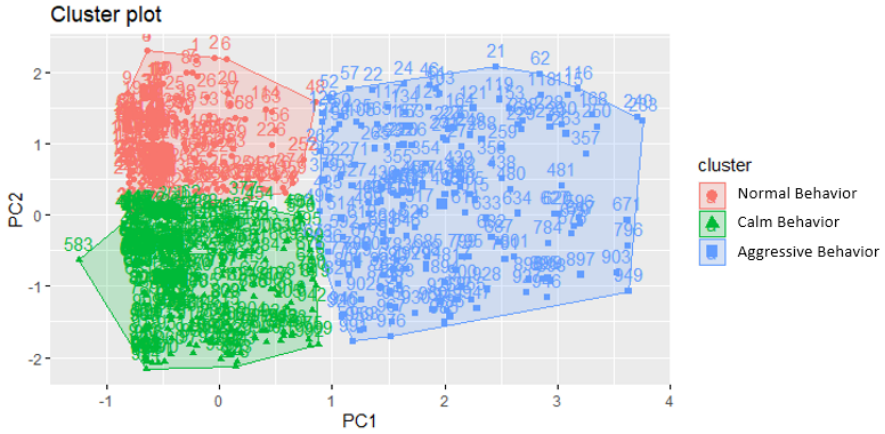

Figure 24. K Means with 3 Clusters on real data

Table 5. K-Means value for 3 Clusters

Cluster	Size	Mean PC1	Mean PC2
Cluster 1 Red	359	-0.8865814	1.3364832
Cluster 2 Green	465	-0.6430742	-1.1096038
Cluster 3 Blue	170	3.6312483	0.2127547

5.3 Comparing different drivers

In this section, the algorithm developed in this master's thesis was utilized to analyze all of the collected data for each of the five vehicles. The aim was to compare the driving styles and levels of aggressiveness among the drivers. The first comparison that was conducted involved determining the duration that each driver spent in the pre-defined zones of behavior. The results for all five vehicles are presented in Table 6.

Table 6. Comparison of aggressiveness of different drivers according to their behavior inside the vehicle

Vehicle	% Aggresive Time	% Normal Time	% Calm Time	% Stopped Time
VW-432	7	19.75	65	8.25
VW-433	30.25	28.25	36.75	4.75
VW-437	14.75	32.25	48.5	4.5
CTV-02	59.25	16.25	3.75	20.75
CTV-04	60.5	16	2.75	20.75

The analysis revealed that there were significant differences in the driving behavior of the five drivers. For instance, drivers of vehicle CTV-02 have a significantly higher score for aggressive driving compared to drivers of vehicle VW-432. This indicates that the driving style of vehicle CTV-02 is more likely to be associated with aggressive behavior on the road.

Another important finding from the analysis was the identification of common patterns in driving behavior across the different vehicles. For instance, three drivers of VW vehicles spent the longest time in the normal driving zone, which suggests that this is the most common mode of driving in that region or type of truck. It is plausible that the three vehicles identified as spending more time in the calm zone have specific applications that do not require aggressive driving behavior. For instance, it is possible that they are used for commercial purposes, such as a garbage truck, which typically involves a slower pace of driving and less frequent acceleration and braking.

An additional comparison of the different vehicles was conducted by applying the K-Means clustering algorithm to all of the collected data. The resulting clusters for each vehicle are depicted in Figures 25, 26, 27, 28, and 29, according to the methodology explained in Section 5.2.2.

By examining the clustering patterns, it is possible to identify similarities and differences in the driving behaviors of the sample drivers. For instance, Figures 27 and 28 show that the driving styles of drivers in two different Volkswagen vehicles were more similar to each other than to the other vehicles in the sample. In contrast, Figures 26 and 27 suggest that the driving styles of the drivers in the two commercial vehicles were more diverse.

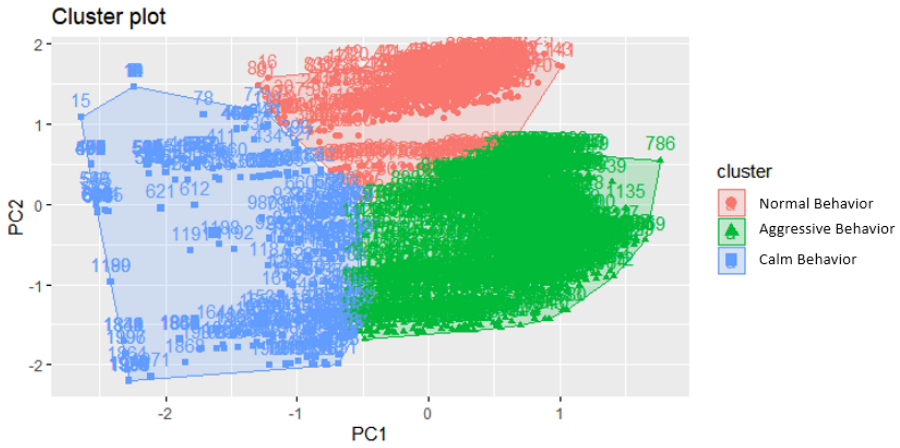

Figure 25. K Means with 3 Clusters on real data for VW 432

Table 7 presents a comparison of the values for all three clusters for the five different vehicles. It is evident that the values for each vehicle are distinct from one another, indicating that each driver has a unique driving style. However, despite the differences, it is possible to observe some common patterns among the driving styles, particularly in terms of aggressiveness. Vehicles labeled as "CTV" refer to medium and heavy-weight models primarily intended for highway driving. When drivers are on highways, they tend to exhibit higher speeds and, consequently, more aggressive behavior, as considered in the proposed algorithm. On the other hand, vehicles marked with the "VW" designation are

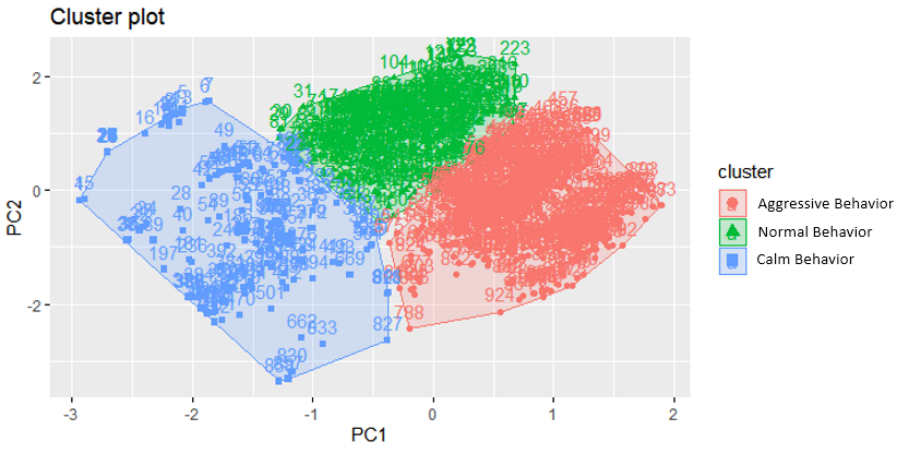

Figure 26. K Means with 3 Clusters on real data for VW 433

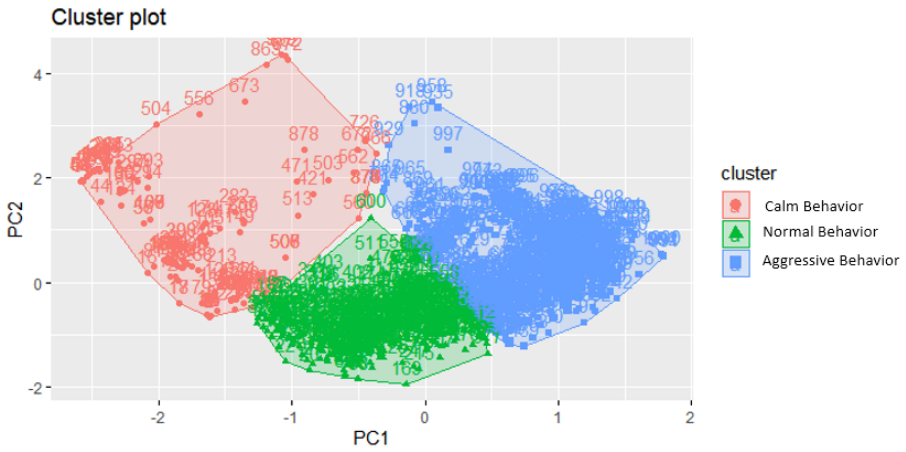

Figure 27. K Means with 3 Clusters on real data for VW 437

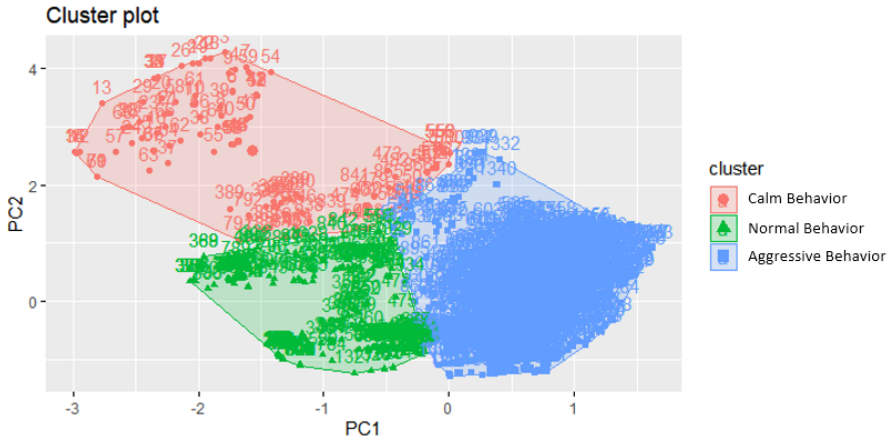

Figure 28. K Means with 3 Clusters on real data for CTV 02

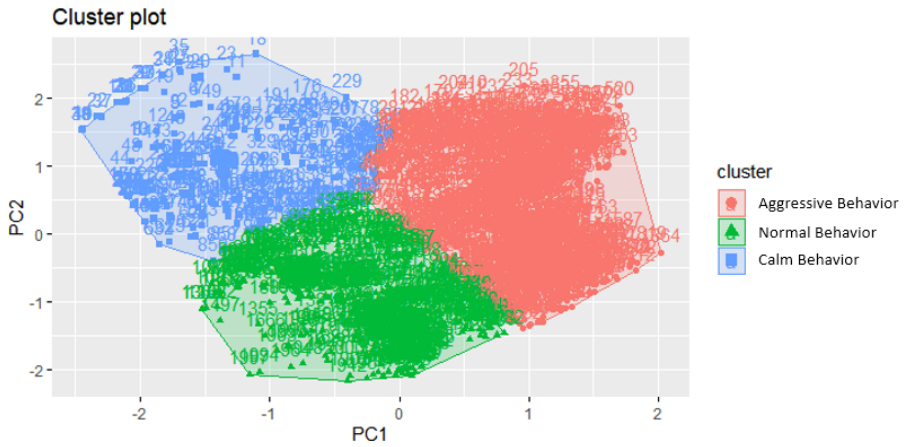

Figure 29. K Means with 3 Clusters on real data for CTV 04

lightweight (Delivery Models) and commonly operate within city limits. In such urban environments, where heavy traffic situations are prevalent, it becomes more challenging to maintain an aggressive driving style.

Table 7. K-Means value comparison between 5 vehicles

Vehicle	Cluster	Size	Mean PC1	Mean PC2
VW-432	Cluster 1 Red	480	0.0574008	1.8331888
VW-432	Cluster 2 Green	1104	1.0486752	-0.6092294
VW-432	Cluster 3 Blue	413	-2.8699512	-0.5020373
VW-433	Cluster 1 Red	490	1.3742172	-0.4096037
VW-433	Cluster 2 Green	308	-0.4562717	1.1301706
VW-433	Cluster 3 Blue	189	-2.8192315	-0.7798238
VW-437	Cluster 1 Red	158	-3.2558757	0.9540605
VW-437	Cluster 2 Green	351	-0.7117164	-0.9829618
VW-437	Cluster 3 Blue	494	1.5470462	0.3932754
CTV-02	Cluster 1 Red	122	-2.957708	3.06484792
CTV-02	Cluster 2 Green	573	-2.203714	-0.74262074
CTV-02	Cluster 3 Blue	1305	1.244114	0.03954807
CTV-04	Cluster 1 Red	931	1.537678	0.4869262
CTV-04	Cluster 2 Green	690	-0.789606	-1.3386783
CTV-04	Cluster 3 Blue	379	-2.339710	1.2410547

Overall, the analysis provides valuable insights into the driving behavior of the sample population and highlights the importance of understanding individual driving styles in promoting safer driving practices.

5.4 Summary

This chapter presents the results obtained from the implementation of two distinct techniques for determining driver behavior. Firstly, a method based on time windows was

applied to identify aggressive driving behavior during specific periods of the trip. This approach was found to be effective when using jerk as a variable, but its performance depends on the chosen window size.

Subsequently, the K-Means technique was employed to group the extracted data into clusters. This unsupervised machine learning method required no prior assumptions about the data and helped to identify distinct patterns of behavior. To reduce the dimensionality of the data, PCA was employed. The results indicated that three distinct groups of data were sufficient to represent the driver's behavior effectively.

Finally, a comparison of the two datasets was performed, highlighting the strengths and weaknesses of each approach.

It is important to notice that this model classify the driver behavior, more specific if the driver is being aggressive during the trip. This model does not classify the driver itself since for this achievement more information should be added to the model, considering that every person has a unique way of driving. The measurements made in 3 happened in different road types and in an uncontrolled environment.

As mentioned in Chapter 3, the methodologies and findings of this study hold applicability across various vehicle classes and geographical contexts. The information utilized in this research is widely available in almost every vehicle, allowing for the potential application of the study's insights in different countries. However, it is crucial to consider the specific legislation and regulations pertaining to driver behavior in each country, as the definition and perception of aggressive driving may vary depending on the road conditions and traffic norms. Furthermore, adapting the predefined limits and parameters for jerk calculation to align with the relevant legislation of the specific country becomes necessary. In the case of this study, the Brazilian legislation for commercial vehicles was considered when defining the thresholds and parameters related to jerk. By taking into account the local legislation and contextual factors, this research can be effectively applied and tailored to different regions and vehicle types, promoting a comprehensive understanding of driver behavior in diverse settings.

6 Conclusion

This master thesis presents a detailed description of the process of information extraction from a vehicular network, analysis of the extracted features, and driver classification based on the extracted data. The previous chapters thoroughly explain the extraction, selection, and classification of data. The driver scoring method used in this study can dynamically identify aggressive driving behavior during pre-defined time windows. This technique uses the calculation of jerk derived from the acquired data. In addition to this approach, another technique was explored in the previous chapter to group different behaviors into data clusters using the K-Means technique.

The main contribution of this research was the extraction, treatment, and analysis of data from vehicles operating in uncontrolled environments with different drivers. In Chapter 2, various techniques were examined, and K-Means was selected due to its advantages in handling large datasets and clustering data effectively. Two key advantages of applying K-Means in such scenarios are scalability and ease of implementation.

Firstly, K-Means demonstrates excellent scalability, enabling efficient processing of large datasets. This characteristic is particularly valuable when dealing with extensive vehicle data involving multiple drivers. Regardless of the dataset size, K-Means can produce reliable and relatively fast results, making it well-suited for analyzing large volumes of vehicle data.

Secondly, K-Means is known for its simplicity and ease of implementation. The algorithm is intuitive, making it accessible to both researchers and practitioners. Its straightforward nature facilitates the clustering of extracted vehicle data, allowing for meaningful comparisons and analysis of different driver behaviors.

Data extraction involved using an electronic control unit (ECU) installed within the vehicles, with the information transmitted to the cloud and subsequently downloaded onto a local computer for analysis using Python and R Studio. During the research, a preliminary analysis of relevant features was conducted, focusing on information that was common across all vehicles to enable effective comparisons.

Another important step in the research was the application of principal component analysis (PCA) to reduce the dimensionality of the data and enhance its visualization. This dimensionality reduction technique proved crucial in applying the unsupervised learning technique of K-Means and visually representing the aggressive driving behavior exhibited by each driver.

While this master's thesis has yielded significant results, it is important to ac-

knowledge and address certain limitations that should be considered in future research. Firstly, the scope of this work only encompassed a limited set of vehicle information, as the primary objective was to compare different vehicles that had minimal shared data. Speed and its related information were the most utilized variables, which aligns with prior research indicating its strong influence on driving aggressiveness. However, if the vehicles were equipped with steering angle sensors, it would have been possible to measure aggressive wheel behavior, thereby enhancing the algorithm.

Furthermore, this work did not incorporate real-time detection of road type or its integration with jerk analysis. This limitation arose from the lack of collected information required to determine such conditions, particularly the absence of GPS data.

Additionally, it is essential to note that this study only considered a limited number of vehicles, all from the Volkswagen brand. Although the developed algorithm could potentially be applied to vehicles of different brands and types, the adaptations necessary for integrating diverse information and sensors from each vehicle were not accounted for within this dissertation. By addressing these limitations in future research, a more comprehensive understanding of driver behavior can be achieved.

The obtained results were deemed satisfactory and have advanced the state-of-the-art of driver behavior classification, being one of the pioneering studies conducted on commercial vehicles in an uncontrolled environment. The final sections below will address the threats to validity that were mitigated in this thesis and provide recommendations for future research.

6.1 Threats to Validity

Validation refers to how effectively a task achieves its intended purpose [74]. To ensure the validity of this dissertation and its potential usefulness in future research, several measures were taken:

Threats to External Validity: This refers to the extent to which the extracted data can be generalized to a larger population beyond the scope of the research. To minimize this risk, real-world data from different drivers and travel days were collected in uncontrolled environments. As a result, this thesis could be extended to other datasets.

Threats to Internal Validity: This refers to the reliability of the researcher's work [75]. One of the significant risks is that the researcher may overlook relevant information or fail to analyze the data appropriately. To mitigate this risk, data extraction was carried out by multiple individuals and verified on an external server.

Threats to Conclusion Validity: This refers to the accuracy of the experimental data and results [75]. To minimize this risk, the data extraction and analysis followed

scientific methodologies and were analyzed accordingly.

6.2 Recommendation for future work

This work made a significant contribution to the analysis of aggressive driver behavior in uncontrolled environments. As for future opportunities for development, the author envisions the following steps:

- In order to enhance the accuracy of the driver behavior classification, it is recommended to develop a driver profile that takes into account a wider range of driver actions along the route. Additionally, it is important to consider the demographic profile of the driver, as this data could potentially influence the classification results based on the theoretical framework studied.
- To optimize the computational cost-benefit of the discussed clustering methods in this thesis, a comparison between them should be carried out.
- To enhance the accuracy of determining a driver's aggressiveness, it is recommended to develop a supervised learning algorithm that can detect the aggressive behavior in real-time during the driving process, rather than only after the trip. Additionally, a system could be developed to provide the driver with recommendations for safe driving practices that can help prevent traffic accidents.
- Analyze the impact of assisted driving technologies, such as autonomous vehicles, on the driver's score.
- Integrate a system with a mobile application to inform drivers about their performance, or notify authorities or fleet owners if a driver is exhibiting aggressive behavior on a highway.

Apêndices

APÊNDICE A – R Studio Code

library(tidyverse)

library(readr)

library(ggbiplot)

library(slider)

library(lubridate)

library(ggpubr)

library(ggnewscale)

library(ggtext)

library(hexbin)

library(corrplot)

library(PCAtools)

library(dplyr)

library(data.table)

library(factoextra)

library(ClusterR)

library(cluster)

$\text{time}_{w}indow = 5$

$\text{time}_{b}egin = 0$

$\text{time}_{f}inal = 10000$

$\text{agg}_{t}hreshold = 1.0$

$\text{calm}_{t}hreshold = 0.5$

$\text{nospeed}_{t}hreshold = 0.0$

$\text{vertical}_{l}ines_{a}gg < -c(0)$

$\text{vertical}_{l}ines_{n}ormal < -c(0)$

$\text{vertical}_{l}ines_{c}alm < -c(0)$

$\text{vertical}_{l}ines_{n}ospeed < -c(0)$

if(1)

$df_VW433 = read_delim(file = "compiladoVW433.csv", ";")$

$df_VW432 = read_delim(file = "compiladoVW432.csv", ";")$

$df_VW437 = read_delim(file = "compiladoVW437.csv", ";")$

$df_CTV02 = read_delim(file = "compiladoCTV02.csv", ";")$

$df_CTV04 = read_delim(file = "compiladoCTV04.csv", ";")$

if(0)

df4 = df %>% select(‘Time Dif‘, ‘Odômetro total (m)‘, ‘Economia de combustível (kml)‘,

‘Velocidade Real (km/h)‘, ‘Velocidade Do Motor (rpm)‘,

‘Pedal De Freio‘, ‘Pedal De Embreagem‘,

‘Posição Do Pedal De Aceleração (%)‘, ‘Motor

Real - % Torque (%)‘)

$df4_without NA < -na.omit(df4)$

colnames(df4) <- c("Time", "$Total_Odometer$", "$Fuel_Economy$", "$Vehicle_Speed$",

"$Engine_Speed$", "$Brake_Pedal$", "$ClutchPedal$",

"Gas_Pedal", "$Engine_Torque$")

if(1)

$df = df_VW432[c(as.integer((time_begin/5)) : as.integer((time_final/5))), c(1 : 23)]$

df4 = df %>% select(‘Time dif‘, ‘Odometro total‘, ‘Economia de combustivel‘,

‘Velocidade Real‘, ‘Velocidade Do Motor‘,

‘Pedal De Freio‘, ‘Pedal De Embreagem‘,

‘Posicao Do Pedal De Aceleracao‘)

$df4_without NA < -na.omit(df4)$

colnames(df4) <- c("Time", "$Total_Odometer$", "$Fuel_Economy$", "$Vehicle_Speed$",

"$Engine_Speed$", "$Brake_Pedal$", "$ClutchPedal$",

"Gas_Pedal")

if(1)

df5 <- df4

df5$Acceleration <- c(NA,with(df5,diff($Vehicle_Speed$)/$diff(Time)$))

df5$Jerk <- c(NA,with(df5,diff(Acceleration)/diff(Time)))

if(1)

df6 <- df5 %>% select(Vehicle$_{S}peed, Time, Acceleration, Jerk$)

GroupLabels <- 0:(nrow(df6) - 1)%/% time$_{w}indow$

df6$Group <- GroupLabels

dt<-data.table(df6)

dt$_{f}$ <$-dt$[, .($min_{t}ime = min(Time)$,

sd$_{j}erk = sd(Jerk)$,

gama = (sd(Jerk)/0.2732),

mean$_{s}peed = mean(Vehicle_{S}peed)$,

sd$_{s}peed = sd(Vehicle_{S}peed)$,

median$_{s}peed = median(Vehicle_{S}peed)$,

mean$_{a}cceleration = mean(Acceleration)$,

sd$_{a}cceleration = sd(Acceleration)$,

median$_{a}cceleration = median(Acceleration)$,

mean$_{j}erk = mean(Jerk)$,

median$_{j}erk = median(Jerk))$,

by = .(Group)]

if(0) df2 = df$_{a}ll$% > %$select(Fuel_{c}onsumption, Engine_{s}peed$,

Vehicle$_{s}peed, Steering_{w}heel_{a}ngle$,

Steering$_{w}heel_{s}peed, Accelerator_{P}edal_{v}alue$,

Engine$_{t}orque$,

Engine$_{c}oolant_{t}emperature$,

‘Acceleration$_{s}peed_{-L}ateral$‘,

‘Acceleration$_{s}peed_{-L}ongitudinal$‘,

Intake$_{a}ir_{p}ressure)$

df$_{a}ll$\$$range = cut(df2$Engine_{s}peed, c(0, 600, 2000, 4000, 6300))$

levels(df$_{a}ll$\$$range) = c($"$Baixa$"$, $"$Ideal$"$, $"$Alta$"$, $"$MuitoAlta$"$)

df3 = df$_{a}ll$% > %$select(X1, Fuel_{c}onsumption, Engine_{s}peed$,

Vehicle$_{s}peed, Steering_{w}heel_{a}ngle$,

Accelerator$_{P}edal_{v}alue, Engine_{t}orque)$

```
if(0)
p1 = ggplot(data = df3) +
geom_point(mapping = aes(x = X1, y = Fuel_consumption))
p2 = ggplot(data = df3) +
geom_point(mapping = aes(x = X1, y = Engine_speed))
p_all = ggarrange(p1, p2, ncol = 1)
print(p_all)
if(0)
corr_variables < −cor(df4_withoutNA)
corrplot(corr_variables, method = "circle", order = "hclust", addrect
= 3,
tl.cex = 0.5, tl.col = "black")
corrplot(corr_variables, order = "AOE", addCoef.col = "gray")
corrplot.mixed(corr_variables, order = "AOE")
if(1)
m = as.matrix(df4_withoutNA)
m.pca = prcomp(m, scale. = T)
summary(m.pca)
pa = ggbiplot(m.pca, obs.scale = 1, var.scale = 1,
ellipse = T,choices = c(1,2)) +
geom_point(aes(color = df1$range))+
theme_minimal()
print(pa)
fviz_eig(m.pca)
if(1)
m = as.matrix(df4_withoutNA)
ms = scale(m)
ms.svd = svd(ms)
if(1)
mt = tibble(x=1:length(ms.svd$d), y=ms.svd$d)
```

p = ggplot(mt) +

geom$_l ine(mapping = aes(x = x, y = y))$

print(p)

mt = as$_t ibble(ms.svd\$u[, 3 : 4])$

p1 = ggplot(mt) +

geom$_p oint(mapping = aes(x = V1, y = V2, color = df1\$range))+$

theme$_m inimal()$

print(p1)

if(1)

p$_p air = pairsplot(pca(ms))$

print(p$_p air)$

if(1)

j = 1;

k = 1;

l = 1;

m = 1;

total$_a gg = 0;$

total$_n ormal = 0;$

total$_c almo = 0;$

total$_p arado = 0;$

loop$_i = 1$

agg$_a nt < -FALSE$

norm$_a nt < -FALSE$

calm$_a nt < -FALSE$

nospeed$_a nt < -FALSE$

for (i in 1:nrow(dt$_f$))

if(is.na(dt$_f[i]\$gama)$)

else if (dt$_f[i]\$gama >= agg_t hreshold)$

total$_a gg = total_a gg + 1$

if (agg$_a nt == FALSE)$

vertical$_l ines_a gg[j] = dt_f[i]\$min_t ime$

j = j+1

agg$_a nt = TRUE$

norm$_a nt = FALSE$

calm$_a nt = FALSE$

nospeed$_a nt = FALSE$

else if (dt$_f[i]\$gama < agg_t hresholddt_f[i]\$gama >= calm_t hreshold$)

total$_n ormal = total_n ormal + 1$

if (norm$_a nt == FALSE$)

vertical$_l ines_n ormal[k] = dt_f[i]\$min_t ime$

k = k+1

agg$_a nt = FALSE$

norm$_a nt = TRUE$

calm$_a nt = FALSE$

nospeed$_a nt = FALSE$

else if(dt$_f[i]\$gama < calm_t hresholddt_f[i]\$gama > nospeed_t hreshold$)

total$_c almo = total_c almo + 1$

if (calm$_a nt == FALSE$)

vertical$_l ines_c alm[l] = dt_f[i]\$min_t ime$

l = l+1

agg$_a nt = FALSE$

norm$_a nt = FALSE$

calm$_a nt = TRUE$

nospeed$_a nt = FALSE$

loop$_i = loop_i + 1$

else if (dt$_f[i]\$mean_s peed == 0$)

total$_p arado = total_p arado + 1$

if (nospeed$_a nt == FALSE$)

vertical$_l ines_n ospeed[m] = dt_f[i]\$min_t ime$

m = m+1

$\text{agg}_a nt = FALSE$

$\text{norm}_a nt = FALSE$

$\text{calm}_a nt = FALSE$

$\text{nospeed}_a nt = TRUE$

if(1)

par(mfrow=c(2,2))

p1 = ggplot(data = df5) +

$\text{geom}_l ine(mapping = aes(x = `Time`, y = `Vehicle_Speed`))$

print(p1)

p2 = ggplot(data = df5) +

$\text{geom}_l ine(mapping = aes(x = `Time`, y = `Acceleration`))$

print(p2)

p3 = ggplot(data = df5) +

$\text{geom}_l ine(mapping = aes(x = `Time`, y = `Jerk`))$

print(p3)

$\text{p}_t este < -p1 + geom_v line(xintercept = vertical_l ines_c alm, color = "green")+$

$\text{geom}_v line(xintercept = vertical_l ines_n ormal, color = "blue")+$

$\text{geom}_v line(xintercept = vertical_l ines_a gg, color = "red")+$

$\text{geom}_v line(xintercept = vertical_l ines_n ospeed, color = "yellow")$

$\text{print}(\text{p}_t este)$

print ("% de agressividade:")

print $((\text{total}_a gg/i) * 100)$

print ("% de normal:")

print $((\text{total}_n ormal/i) * 100)$

print ("% de calmo:")

print $((\text{total}_c almo/i) * 100)$

print ("% parado:")

print $((\text{total}_p arado/i) * 100)$

if (1)

$\text{df}_k m = df4\% > \%select(`Vehicle_Speed`, `Gas_Pedal`)$

$\text{dfkm}_{w}ithoutNA < -na.omit(df_{k}m)$

$\text{pca}_{t}ransform = as.data.frame(m.pca\$x[, 1:2])$

set.seed(240) Setting seed

wss <- function(k)

$\text{kmeans}(\text{pca}_{t}ransform, k, nstart = 25)\$tot.withinss$

k.values <- 1:15

$\text{wss}_{v}alues < -map_{d}bl(k.values, wss)$

plot(k.values, $\text{wss}_{v}alues$,

type="b", pch = 19, frame = FALSE,

xlab="Number of clusters K",

ylab="Total within-clusters sum of squares")

$\text{km.res} <- \text{kmeans}(\text{pca}_{t}ransform, 3, iter.max = 10, nstart = 25)$

print(km.res)

$\text{fviz}_{c}luster(km.res, data = pca_{t}ransform)$

Anexos

ANEXO A – Published article

It was published the article "Data Analysis Techniques in Vehicle Communication Networks: Systematic Mapping of Literature" in IEEE Access, Volume 8, 2020. DOI: 10.1109/ACCESS.2020.3034588.

Article Abstract:

Vehicles are becoming more intelligent and connected due to the demand for faster, efficient, and safer transportation. For this transformation, it was necessary to increase the amount of data transferred between electronic modules in the vehicular network since it is vital for an intelligent system's decision-making process. Hundreds of messages travel all the time in a vehicle, creating opportunities for analysis and development of new functions to assist the driver's decision. Given this scenario, this article presents the results of research to found out which data analysis techniques in vehicular communication networks and for which purposes they are designed. The research method adopted was the systematic mapping of literature, where 196 articles were found using a search protocol. All papers were classified according to the established inclusion and exclusion criteria, and the main results contained were discussed. To obtain a clear view of the generated information and support the identification of possible gaps in this field, correlation graphs, and a systematic map was developed. It was possible to verify that the identification of the driver's profile was the most studied application, with the use of neural network techniques to correlate the gathered data.

Bibliography

1 ACIDENTES rodoviários com caminhões. Disponível em: <https://cnt.org.br/acidentes-rodoviarios-caminhoes>. 7, 15, 16, 17, 20

2 VOLKSWAGEN Truck Bus Website. 2023. [Online; accessed 12-April-2023]. Disponível em: <https://www.vwco.com.br/>. 7, 18

3 Robert Bosch GmbH. *CAN-Bus Specifications*. Postfach 50, D-7000, Stuttgart 1, date. This reference does not have a specific publication date. Please replace "date" with the appropriate year of publication. 7, 24

4 NARAYANAN, A.; SIRAVURU, A.; DARIUSH, B. Gated recurrent fusion to learn driving behavior from temporal multimodal data. *IEEE Robotics and Automation Letters*, v. 5, n. 2, p. 1287–1294, 2020. ISSN 23773766 (ISSN). Disponível em: <https://doi.org/10.1109/LRA.2020.2967738>. 7, 26

5 ORTIZ, M. G. *Prediction of Driver Behavior*. Tese (Doktor-Ingenieur (Dr.-Ing.)) — Universität Bielefeld, Technische Fakultät, Universitätsstr. 25, 33615 Bielefeld, Germany, 2014. 7, 19, 20, 27, 28, 29

6 MURPHEY, Y. L.; MILTON, R.; KILIARIS, L. Driver's style classification using jerk analysis. In: IEEE. *2009 IEEE International Conference on Robotics and Automation*. [S.l.], 2009. p. 2822–2827. 9, 32, 56, 59, 60, 61

7 BURTON, A. et al. Driver identification and authentication with active behavior modeling. In: *Int. Conf. Netw. Serv. Manag., CNSM Workshops, Int. Workshop Manag. SDN NFV, ManSDN/NFV Int. Workshop Green ICT Smart Netw., GISN*. [s.n.], 2017. p. 388–393. ISBN 9783901882852 (ISBN). 2016 12th International Conference on Network and Service Management, CNSM 2016 and Workshops, 3rd International Workshop on Management of SDN and NFV, ManSDN/NFV 2016, and International Workshop on Green ICT and Smart Networking, GISN 2016. Disponível em: <https://doi.org/10.1109/CNSM.2016.7818453>. 14

8 FUGIGLANDO, U. et al. Driving behavior analysis through can bus data in an uncontrolled environment. *IEEE Transactions on Intelligent Transportation Systems*, v. 20, n. 2, p. 737–748, 2019. ISSN 15249050 (ISSN). Disponível em: <https://doi.org/10.1109/TITS.2018.2836308>. 14, 30

9 JEONG, Y. et al. An integrated self-diagnosis system for an autonomous vehicle based on an iot gateway and deep learning. *Applied Sciences (Switzerland)*, v. 8, n. 7, 2018. ISSN 20763417 (ISSN). Disponível em: <https://doi.org/10.3390/app8071164>. 14

10 HALIM, Z.; KALSOOM, R.; BAIG, A. R. Profiling drivers based on driver dependent vehicle driving features. *Applied Intelligence*, Springer Science and Business Media LLC, v. 44, n. 3, p. 645–664, nov. 2015. Disponível em: <https://doi.org/10.1007/s10489-015-0722-6>. 16

11 WANG, Y. et al. The validity of driving simulation for assessing differences between in-vehicle informational interfaces: A comparison with field testing. *Ergonomics*, v. 53, n. 3, p. 404–420, 2010. 17

12 GODLEY, S. T.; TRIGGS, T. J.; FILDES, B. N. Driving simulator validation for speed research. *Accident Analysis & Prevention*, v. 34, n. 5, p. 589–600, 2002. 17

13 EZZINI, S.; BERRADA, I.; GHOGHO, M. Who is behind the wheel? driver identification and fingerprinting. *Journal of Big Data*, v. 5, n. 1, 2018. ISSN 21961115 (ISSN). Disponível em: <https://doi.org/10.1186/s40537-018-0118-7>. 18

14 THRUN, S. *What We're Driving At*. 2010. <http://googleblog.blogspot.fr/2010/10/what-were-driving-at.html>. Accessed: 2023-03-12. 19

15 ABUKHALIL, T. et al. Fuel consumption using obd-ii and support vector machine model. *Journal of Robotics*, v. 2020, 2020. ISSN 16879600 (ISSN). Disponível em: <https://doi.org/10.1155/2020/9450178>.

16 AVATEFIPOUR, O. et al. An intelligent secured framework for cyberattack detection in electric vehicles' can bus using machine learning. *IEEE Access*, v. 7, p. 127580–127592, 2019. ISSN 21693536 (ISSN). Disponível em: <https://doi.org/10.1109/ACCESS.2019.2937576>.

17 CHEN, Z. et al. Multi-dimensional and multi-scale modeling of traffic state in jiangxi expressway based on vehicle network. *International Journal of Performability Engineering*, v. 15, n. 12, p. 3287–3294, 2019. ISSN 09731318 (ISSN).

18 DELNEVO, G. et al. On combining big data and machine learning to support eco-driving behaviours. *Journal of Big Data*, v. 6, n. 1, 2019. ISSN 21961115 (ISSN). Disponível em: <https://doi.org/10.1186/s40537-019-0226-z>.

19 HANSELMANN, M. et al. Canet: An unsupervised intrusion detection system for high dimensional can bus data. *IEEE Access*, v. 8, p. 58194–58205, 2020. ISSN 21693536 (ISSN). Disponível em: <https://doi.org/10.1109/ACCESS.2020.2982544>.

20 LEE, S. H.; LEE, S.; KIM, M. H. Development of a driving behavior-based collision warning system using a neural network. *International Journal of Automotive Technology*, v. 19, n. 5, p. 837–844, 2018. ISSN 12299138 (ISSN). Disponível em: <https://doi.org/10.1007/s12239-018-0080-6>.

21 LIN, N. et al. An overview on study of identification of driver behavior characteristics for automotive control. *Mathematical Problems in Engineering*, v. 2014, 2014. ISSN 1024123X (ISSN). Disponível em: <https://doi.org/10.1155/2014/569109>.

22 PARK, J. et al. Road surface classification using a deep ensemble network with sensor feature selection. *Sensors (Switzerland)*, v. 18, n. 12, 2018. ISSN 14248220 (ISSN). Disponível em: <https://doi.org/10.3390/s18124342>.

23 PARK, S.; CHOI, J.-Y. Malware detection in self-driving vehicles using machine learning algorithms. *Journal of Advanced Transportation*, v. 2020, 2020. ISSN 01976729 (ISSN). Disponível em: <https://doi.org/10.1155/2020/3035741>.

24 SUN, Y. et al. Research on safe driving behavior of transportation vehicles based on vehicle network data mining. *Transactions on Emerging Telecommunications Technologies*, v. 31, n. 5, 2020. ISSN 21615748 (ISSN). Disponível em: <https://doi.org/10.1002/ett.3772>.

25 TUMAS, P.; NOWOSIELSKI, A.; SERACKIS, A. Pedestrian detection in severe weather conditions. *IEEE Access*, v. 8, p. 62775–62784, 2020. ISSN 21693536 (ISSN). Disponível em: <https://doi.org/10.1109/ACCESS.2020.2982539>.

26 WANG, H. et al. A driver's car-following behavior prediction model based on multi-sensors data. *Eurasip Journal on Wireless Communications and Networking*, v. 2020, n. 1, 2020. ISSN 16871472 (ISSN). Disponível em: <https://doi.org/10.1186/s13638-020-1639-2>.

27 XIAO, J.; WU, H.; LI, X. Internet of things meets vehicles: Sheltering in-vehicle network through lightweight machine learning. *Symmetry*, v. 11, n. 11, 2019. ISSN 20738994 (ISSN). Disponível em: <https://doi.org/10.3390/sym11111388>.

28 YAN, F. et al. Driving style recognition based on electroencephalography data from a simulated driving experiment. *Frontiers in Psychology*, v. 10, n. MAY, 2019. ISSN 16641078 (ISSN). Disponível em: <https://doi.org/10.3389/fpsyg.2019.01254>.

29 ZARDOSHT, M.; BEAUCHEMIN, S. S.; BAUER, M. A. Identifying driver behavior in preturning maneuvers using in-vehicle canbus signals. *Journal of Advanced Transportation*, v. 2018, 2018. ISSN 01976729 (ISSN). Disponível em: <https://doi.org/10.1155/2018/5020648>.

30 ZHANG, J. et al. A deep learning framework for driving behavior identification on in-vehicle can-bus sensor data. *Sensors (Switzerland)*, v. 19, n. 6, 2019. ISSN 14248220 (ISSN). Disponível em: <https://doi.org/10.3390/s19061356>.

31 ZHOU, A.; LI, Z.; SHEN, Y. Anomaly detection of can bus messages using a deep neural network for autonomous vehicles. *Applied Sciences (Switzerland)*, v. 9, n. 15, 2019. ISSN 20763417 (ISSN). Disponível em: <https://doi.org/10.3390/app9153174>.

32 NARAYANAN, S. N.; MITTAL, S.; JOSHI, A. Using data analytics to detect anomalous states in vehicles. *Department of Computer Science & Electrical Engineering, University of Maryland Baltimore County*, Baltimore, 21227, Maryland, U.S.A., 2015. 25

33 XIAOLIANG, M. A neural-fuzzy framework for modeling car-following behavior. In: *Systems, Man and Cybernetics*. [S.l.: s.n.], 2006. p. 1178–1183. 27

34 CHANDLER, R. E.; HERMAN, R.; MONTROLL, E. W. Traffic dynamics: Studies in car following. *Operations Research*, v. 6, n. 2, p. 165–184, 1958. 28

35 TREAT, J. R. et al. *Tri-Level Study of the Causes of Traffic Accidents: Final Report*. Springfield, VA, 1979. 28

36 MILLER, G.; BEN-ARI, O. T. Driving styles among young novice drivers–the contribution of parental driving styles and personal characteristics. *Accident Analysis & Prevention*, v. 42, n. 2, p. 558–570, 2010. 29

37 KIM, K. et al. Personal and behavioral predictors of automobile crash and injury severity. *Accident Analysis &amp Prevention*, Elsevier BV, v. 27, n. 4, p. 469–481, ago. 1995. Disponível em: <https://doi.org/10.1016/0001-4575(95)00001-g>. 29

38 SCOTT-PARKER, B. et al. Speeding by young novice drivers: What can personal characteristics and psychosocial theory add to our understanding? *Accident Analysis & Prevention*, v. 50, p. 242–250, 2013. 29

39 CHOUDHARY, P.; VELAGA, N. R. Mobile phone use during driving: Effects on speed and effectiveness of driver compensatory behaviour. *Accident Analysis &amp Prevention*, Elsevier BV, v. 106, p. 370–378, set. 2017. Disponível em: <https://doi.org/10.1016/j.aap.2017.06.021>. 29

40 HECK, K. E.; CARLOS, R. M. Passenger distractions among adolescent drivers. *Journal of Safety Research*, Elsevier BV, v. 39, n. 4, p. 437–443, jan. 2008. Disponível em: <https://doi.org/10.1016/j.jsr.2008.03.003>. 29

41 AARTS, L.; SCHAGEN, I. v. Driving speed and the risk of road crashes: A review. *Accident Analysis & Prevention*, Elsevier, v. 38, p. 215–224, 2006. 30

42 CONSTANTINESCU, Z.; MARINOIU, C.; VLADOIU, M. Driving style analysis using data mining techniques. *International Journal of Computers Communications & Control*, v. 5, n. 5, p. 654–663, 2010. 30

43 PAPAIOANNOU, P. Driver behaviour, dilemma zone and safety effects at urban signalised intersections in greece. *Accident Analysis & Prevention*, Elsevier, v. 39, n. 1, p. 147–158, 2007. 31

44 KEDAR-DONGARKAR, G.; DAS, M. Driver classification for optimization of energy usage in a vehicle. *Procedia Computer Science*, Elsevier, v. 8, p. 388–393, 2012. 31

45 AUGUSTYNOWICZ, A. Preliminary classification of driving style with objective rank method. *International journal of automotive technology*, v. 10, n. 5, p. 607–610, 2009. 31

46 GE, A. et al. Study on automobile intelligent shift architecture. *China Mechanical Engineering*, v. 5, n. 18, p. 106–109, 2001. 31

47 JOHNSON, D. A.; TRIVEDI, M. M. Driving style recognition using a smartphone as a sensor platform. In: IEEE. *Intelligent Transportation Systems (ITSC), 2011 14th International IEEE Conference on*. [S.l.], 2011. p. 1609–1615. 31

48 DOSHI, A.; TRIVEDI, M. M. Examining the impact of driving style on the predictability and responsiveness of the driver: Real-world and simulator analysis. In: IEEE. *Intelligent Vehicles Symposium (IV), 2010 IEEE*. [S.l.], 2010. p. 232–237. 31

49 RYGULA, A. Driving style identification method based on speed graph analysis. In: *International Conference on Biometrics and Kansei Engineering*. [S.l.: s.n.], 2009. p. 76–79. 31

50 LIU, Y. et al. Understanding of internal clustering validation measures. In: *2010 IEEE International Conference on Data Mining*. [S.l.: s.n.], 2010. p. 911–916. 32

51 KALSOOM, R.; HALIM, Z. Clustering the driving features based on data streams. In: IEEE. *INMIC*. [S.l.], 2013. p. 89–94. 32

52 LANGARI, R.; WON, J.-S. Intelligent energy management agent for a parallel hybrid vehicle-part i: system architecture and design of the driving situation identification process. *IEEE Transactions on Vehicular Technology*, IEEE, v. 54, n. 3, p. 925–934, 2005. 32

53 AASTHA, J.; RAJNEET, K. Review: Comparative study of various clustering techniques in data mining. *International Journal of Advanced Research in Computer Science and Software Engineering*, v. 3, p. 55–57, March 2013. 33, 34

54 NISHA; KAUR, P. J. Cluster quality based performance evaluation of hierarchical clustering method. In: IEEE. *2015 1st International Conference on Next Generation Computing Technologies (NGCT)*. [S.l.], 2015. p. 649–653. 33

55 WAHIDAH, H.; PEY, L. et al. Application of data mining techniques for improving software engineering. In: *The 5th International Conference on Information Technology*. [S.l.: s.n.], 2011. v. 2, p. 1–5. 33

56 KUSETOGULLARI, H. Unsupervised text binarization in handwritten historical documents using k-means clustering. In: SPRINGER. *Proceedings of SAI Intelligent Systems Conference (IntelliSys) 2016*. [S.l.], 2018. v. 16, p. 65–75. 33

57 HENRIQUE, R. R.; AHMED, E. A. A. Proposed application of data mining technique for clustering software projects. *INFOCOMP- special edition*, p. 43–48, Jul 2010. 33

58 KRIEGEL, H.-P.; KRÖGER, P.; ZIMEK, A. Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, v. 3, n. 1, p. 1, 2009. 33

59 YAN, J. et al. Effective and effcient dimensionality reduction for large-scale and streaming data preprocessing. *IEEE Transactions on Knowledge and Data Engineering*, IEEE, v. 18, n. 3, p. 320–333, 2006. 33, 34

60 GHEYAS, I. A.; SMITH, L. S. Feature subset selection in large dimensionality domains. *Pattern Recognition*, Elsevier, v. 43, n. 1, p. 5–13, 2010. ISSN 0031-3203. 34

61 CONTRIBUTORS, W. *Dimensionality reduction — Wikipedia, The Free Encyclopedia*. 2019. Accessed: April 19, 2019. Disponível em: <https://en.wikipedia.org/wiki/Dimensionality_reduction>. 34

62 ABDI, H.; WILLIAMS, L. J. Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, Wiley Online Library, v. 2, n. 4, p. 433–459, 2010. 34, 35

63 MAATEN, L. v. d.; HINTON, G. Visualizing data using t-sne. *Journal of machine learning research*, v. 9, n. Nov, p. 2579–2605, 2008. 36

64 KURITA, A. K. *Paper Dissected: 'Visualizing Data using t-SNE' Explained*. 2018. <http://mlexplained.com/2018/09/14/paper-dissected-visualizing-datausing-t-sne-explained/>. 36

65 Reddy, N. R. *Driving Behaviour Classification : An Eco-driving Approach*. 2019. Disponível em: <http://essay.utwente.nl/80158/>. 37, 41, 49

66 JIN, X.; HAN, J. K-means clustering. In: SAMMUT, C.; WEBB, G. I. (Ed.). *Encyclopedia of Machine Learning*. [S.l.]: Springer, 2010. 38

67 PAKHIRA, M. K. A linear time-complexity k-means algorithm using cluster shifting. In: IEEE. *2014 International Conference on Computational Intelligence and Communication Networks*. [S.l.], 2014. p. 1047–1051. 39

68 MURTAGH, F.; CONTRERAS, P. Methods of hierarchical clustering. *CoRR*, abs/1105.0121, 2011. Disponível em: <http://arxiv.org/abs/1105.0121>. 39

69 ESTER, M. et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In: *Kdd*. [S.l.: s.n.], 1996. v. 96, n. 34, p. 226–231. 41

70 DATA acquisition — Wikipedia, the free encyclopedia. 2014. [Online; accessed 12-March-2023]. Disponível em: <https://en.wikipedia.org/w/index.php?title=Data_acquisition&oldid=628710638>. 45

71 DAVIS, G. A.; DAVULURI, S.; PEI, J. P. *A Case Control Study of Speed and Crash Risk, Technical Report 3: Speed as a Risk Factor in Run-off Road Crashes*. University of Minnesota Digital Conservancy, 2006. Disponível em: <https://hdl.handle.net/11299/542>. 54

72 FARMER, C. M. Relationship of traffic fatality rates to maximum state speed limits. *Traffic Injury Prevention*, Taylor & Francis, v. 18, n. 4, p. 375–380, 2017. 54

73 HWANG, C.-p. et al. Apply scikit-learn in python to analyze driver behavior based on obd data. In: IEEE. *2018 32nd International Conference on Advanced Information Networking and Applications Workshops (WAINA)*. [S.l.], 2018. p. 636–639. 55

74 SERBEC, I.; STRNAD, M.; RUGELJ, J. Assessment of wiki-supported collaborative learning in higher education. In: *IEEE*. [S.l.: s.n.], 2010. 73

75 KITCHENHAM, B. A.; BUDGEN, D.; BRERETON, P. *Evidence-based software engineering and systematic reviews*. [S.l.]: CRC press, 2015. v. 4. 73