MASTER THESIS

Driving Behaviour Classification : An Eco-driving Approach

Student :
Navin Ramesh Reddy

Committee University of Twente :
Dr. N. Meratnia
Prof. Dr. P.J.M. Havinga
Ir. E. Molenkamp

Faculty of Electrical Engineering, Mathematics and Computer Science

Pervasive Systems

UNIVERSITY OF TWENTE.

UNIVERSITY OF TWENTE

MASTER THESIS

Driving Behaviour Classification: An Eco-driving Approach

Author:
Navin Ramesh Reddy

Committee:
Dr. N. Meratnia
Prof. Dr. P.J.M. Havinga
Ir. E. Molenkamp

Pervasive Systems
Faculty Of Electrical Engineering, Mathematics And Computer Science

November 25, 2019

Abstract

Driving behavior plays a vital role in determining road safety and also greatly impacts fuel efficiency. Eco-driving is an efficient and economical way of driving, which contributes to the decrease in fuel consumption and pollution. This thesis work deals with the driving behaviour analysis from the perspective of eco-driving rules. Classification of the driving behavior is based on the features which are extracted from the time-series signals collected from On-Board Diagnostic (OBD-II) port of a vehicle. Two methods of classification are proposed: a scoring algorithm based on fuzzy logic and unsupervised learning method. The scoring algorithm is designed to provide a quantitative feedback. The unsupervised learning methods are explored for classifying the drivers behaviour. In order to evaluate these methods, the real-world driving data collected from different vehicles is used. The results show that there is a high correlation between the calculated score and the fuel consumption. Further, unsupervised learning concepts are also employed to distinguish among different driving behaviors.

Acknowledgements

This research is the product of collective efforts put in by many people and I take this opportunity to acknowledge their contributions. First and foremost, I would like to thank my daily supervisor Dr. Nirvana Meratnia for all the guidance and help to me this project would not have been possible; for all the interesting solutions for the problems I faced during work and all the encouragement that pushed me forward to deliver my best.

I would also like to thank my committee members Prof. Dr. Paul Havinga and Ir. E. Molenkamp for their valuable time. Furthermore, I thank my brother Nitin and Dr. Ir. Kyle Zhang for helping me collect data necessary for this master thesis. I would like to thank the Pervasive Systems group members for their wonderful company that made my time during the thesis easier and truly memorable.

I truly acknowledge and thank the secretary Ms. Nicole Baveld for the technical support and smooth organisation through the course of the thesis. At last, I would like to express my hearty gratitude to my parents, family and all my friends for their unwavering faith in me and undying support that kept me strong emotionally through the entire journey of my graduate program.

Contents

List of Figures

List of Tables

List of Abbreviations

AFR	Air-to-Fuel Ratio
COG	Centre of Gravity
CVT	Continuous Variable Transmission
ECU	Engine Control Unit
FIS	Fuzzy Inference System
FD	Fuel Density
FoM	Figure of Merits
GPS	Global Positioning System
MAE	Mean Absolute Error
MAF	Mass Air Flow
MAPE	Mean Absolute Percentage Error
OBD-II	On-board Diagnostics-II
PCA	Principal Component Analysis
PKE	Positive Kinetic Energy
RPA	Relative Positive Acceleration
RPM	Revolutions Per Minute
TPS	Throttle Position Sensor
t-SNE	t-Stochastic Neighbour Embedding
WHO	World Health Organization
WCSS	Within-Cluster Sum of Squares

Chapter 1

Introduction

Driving behavior is an important factor in determining the road safety and its impact on the environment. According to the surveys conducted by different institutions and the literature referred, road accidents stand at 9th position in the list of leading causes of deaths in the world [1]. According to the World Report on Road Traffic Injury and Prevention published by World Health Organization (WHO), nearly 1.25 million people die in road crashes each year and an additional 20-25 million are injured or disabled [2]. Also, according to this report [2], one of the main cause of road accidents is aggressive driving. Aggressive driving is defined by the driving behavior where events such as sudden acceleration, abrupt lane change, harsh braking are present. It is also observed that the rate of accidents is directly linked to the driving behavior where these driving maneuvers are present. Sometimes these maneuvers are inevitable and depends on the driving conditions present on the road. Considering all the cases discussed, it is safe to say that there is a need to monitor the driving behavior. A system is required which is capable of analyzing the driving behavior and providing feedback to cope up with the increasing rate of road accidents.

The automotive industry is largely made up of gasoline and diesel based vehicles. Technological advancements in the past few decades have introduced us to an alternate range of vehicles mostly comprising of electric and hybrid vehicles. Yet the majority of the vehicles active on the roads have internal combustion engines that use gasoline and diesel. Emissions caused due to these gasoline and diesel based vehicles are the primary source of pollution [3]. Transport sector, the fastest growing industry sector among others, is also based on the vehicles with internal combustion engines. As the fastest growing industry sector, its contribution to the climate change is very dominant and it is increasing rapidly [4]. CO_2 and short-lived climate pollutants such as black carbon are the highest contributor for the global warming effect.

To cope up with this ever increasing rate of global warming and climate change, vehicle manufacturers are required to follow various rules and regulations. Cleaner fuels are also introduced to tackle with these environmental changes. It is not possible to completely omit gasoline and diesel based vehicles, but as an individual, it is possible to make efficient use of the available resources and contribute towards reducing the pollution. Following greener driving style, sometimes also referred as "eco-driving" style can lead to the reduction of pollution. Ecological (Eco) driving not only contributes in reducing the pollution but has other benefits too. These benefits are reduced cost of journey and lesser mechanical stress on the engines [5]. A real-time or an offline analysis of the journeys keeping the eco-driving context in picture can be used to improve the driving styles.

Applications of driver behavior analysis extend beyond the safety and ecological factors. Insurance companies are often found providing incentives based on the driver's behavior. A driver with good driving skills and safer driving approach gets more discount on the insurance policy when compared to a driver with aggressive driving skills. Another application of driver behavior analysis can be to track the performance and efficiency of the driver in the logistics sector. Goods and commodities that are being shipped are at a greater risk due to aggressive driving. This risk can be minimized by analyzing the trips made and providing feedback or necessary driving lessons to the driver. This application of driving behavior analysis often comes under fleet management.

Fleet management is defined as the organisation and management of commercial vehicles such as cars and trucks. The main objectives of fleet management are to reduce costs, improve efficiency and ensure compliance of the regulations across the entire fleet. Fuel alone accounts for 33% of the fleet's operating costs [6]. To improve the fuel efficiency, companies invest in training the driver by gathering the data from the trips driven by them. Based on the trip's data, feedback is provided to the driver for improving the efficiency. Surveys conducted in [7] shows that the feedback given to the drivers are qualitative in nature, and lacks quantitative suggestions. Qualitative suggestions are based on the observations made on the driving behaviour such as "efficient" or "inefficient". Quantitative suggestions generally refer to a value, score or a rating. One of the reasons for the feedback to be qualitative is the excessive amount of processing involved in quantitative analysis to generate a score or rating. This excessive processing is present due to the large number of parameters that describe the driver's profile.

The driver's profile comprising of different parameters is gathered from the trips driven by the drivers. This profile accounts for large quantities of information and is generally referred to as Big-Data. To analyze this data and differentiate the driving behaviour, unsupervised learning can be applied. This unsupervised learning is a type of machine learning algorithm which is used to discover the unknown properties of the data. With the application of unsupervised learning on driver's profile, it is possible to classify the driving behavior as "eco" or "aggressive". In the sources and literature reviewed during this thesis work, this method for classification has not been applied for eco-driving classification and will be presented in this work. Apart from this method of classification, fuzzy logic is also used to provide quantitative rating to the drivers.

As per [7], it is evident that providing quantitative or qualitative feedback to the drivers have improved their fuel consumption. This concept is well known as gamification where scoring, competition, and rewards are used to motivate a user. Research has shown that providing feedback through game elements has a better effect compared to the normal advices [8]. To quantify the driver's trip, scoring concept from the gamification can be used.

1.1 Research Questions

In the past, eco-driving training has been given to the drivers, and, on an average a fuel consumption improved by 5-10% [9][10][11]. The literature survey based on driver's behavior (presented in Chapter 2) indicates that there is no eco-driving scoring model available currently that reflects the driving behaviour, whereas, there are several scoring models for aggressive driving [12] [13] [14] [15]. In this thesis work, an approach is taken towards designing a scoring model from the eco-driving rules. Unsupervised learning algorithms have been used to classify whether a driving behaviour is safe or unsafe [16]. Eco-driving is in general attributed to safe driving, however, far too little attention has been paid to shown the correlation between eco-driving and safe driving.

This research work seeks to address the following questions:

- What are the parameters that best characterize the eco-driving behaviour?
- How can a scoring model be developed from these parameters?
- To what extent unsupervised machine learning algorithms are applicable to classify the driving behaviour?
- Are eco-driving and safe driving correlated?
- How do the external factors influence eco-driving?

1.2 Thesis Outline

Chapter 2 provides the background for eco-driving, related work and the basic theory of the methods used. Chapter 3 explains the methods used. Chapter 4 presents the results of the methods. Finally, Chapter 5 concludes the thesis and presents a section for the future work.

Chapter 2

Background

2.1 Eco-driving [17]

It is clearly evident from the recent advances that the world is moving towards economical and ecological technology for transportation. Although engineers and scientists put numerous efforts on developing an efficient engine, when it comes to achieving that efficiency, the highest responsibility lies on the user (driver). Every year the average fuel efficiency of the new vehicles goes up significantly. If a driver is not well educated in terms of eco-driving, all the inputs that goes into making an efficient vehicle is of no use. Therefore, there is a need to educate people about the approach towards eco-driving.

There are 5 golden rules for eco-driving:

1. **Greater anticipation** - Driving by anticipating the road and traffic conditions continuously helps in achieving a greater efficiency since there won't be any abrupt changes to the acceleration or deceleration of the vehicle which is a major cause for loss in efficiency. Anticipating the traffic and road ahead gives a cushion and preparedness and use of vehicles momentum to save the fuel by braking smoothly.

2. **Drive at a constant speed** - The value of fuel efficient speed varies from vehicle to vehicle since it majorly depends on the engine capacity. Therefore, every vehicle has a different cruising speed and it usually is between 60 to 90 km/hr. The best efficiency is achieved when the throttle inputs are minimum and when the vehicle is mostly travelling because of the momentum.

3. **Maintaining optimum air pressure** - Tires are the final point of contact for the vehicle on road, lower tire pressure leads to more resistance which leads to more load on the engine finally decreasing the efficiency. Higher pressures are also unsafe since there is compromise of grip on the road. Optimum tire pressure leads to a low rolling resistance translating to higher fuel efficiency.

4. **Shifting gears at lower engine speeds** - In case of vehicles with gears shifting early would be useful to conserve fuel. Shifting gears up early drops the engine speed considerably and also reduces fuel consumption.

5. **Reducing the unnecessary load in the vehicle** - Any unnecessary weight in the vehicle would bear an extra load on the vehicle which would require extra fuel to move the load hence it is a good idea to remove the ancillary load from the vehicle. While driving at high speeds, open windows create a drag which also bears an extra load on the engine.

2.2 Driving Parameters and Features

Driving behaviour analysis is a complex mechanism. Several features/parameters contribute to identify the driving behaviour as events. In this section, the features that affect the driving behaviour in general will be discussed.

2.2.1 Speed

Speed is an important factor in assessing a drivers performance and it is defined as rate of change of distance per unit time, measured by an internal speedometer with units such as kilometers per hour (km/hr), miles per hour (mph) and meter per second (m/s). It is obtained externally from an On-board diagnostics port-II (OBD-II) or through a GPS device such as smartphones. Aggressive driving is a behavioural pattern in which the driver is associated with over-speeding, swerving and cornering. Eco-driving is a classification in which the driver tends to conserve as much fuel as possible by driving at an economical speed.

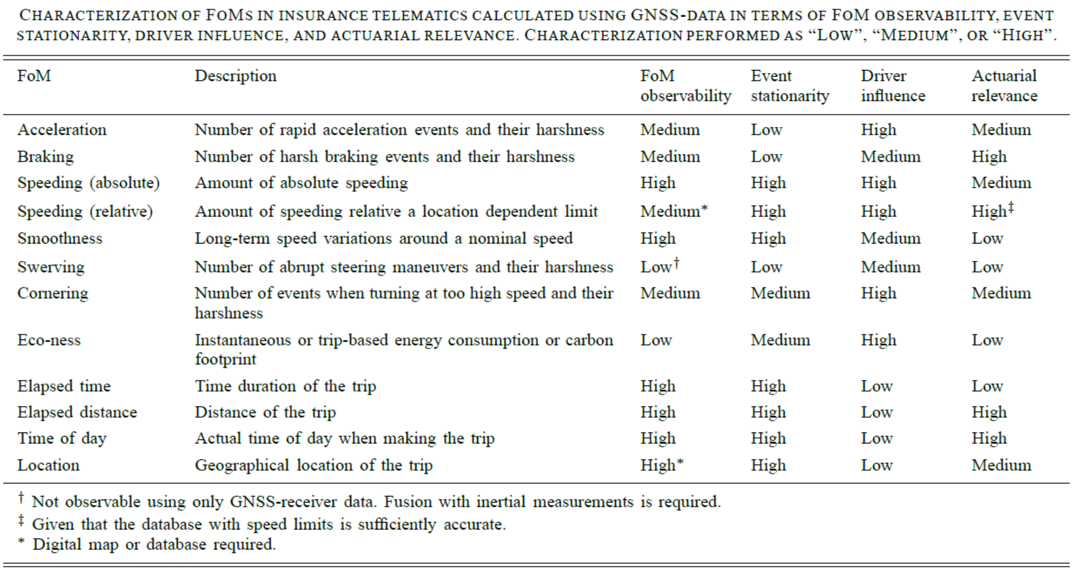
2.2.1.1 Influence of Speed on Aggressive Driving

Risk of accidents increases at higher speeds as found in [18]. Increase in the legal speed limits also caused a higher fatality risk [19]. Smartphone based insurance telematics are disrupting the industry due to the ease of implementation and availability. Handel et al. have listed Figure of Merits (FoMs) that characterize the aggressive nature of the driver using the smartphone GPS and sensors such as accelerometer and gyroscope [13]. These FoM's shown in Figure 2.1 are measured based on the following -

- FoM observability - The correlation between the sensor and measurements.
- Event stationary - The time length during which the events make up for the FoM.
- Actuarial relevance - The importance of the FoM for the risk assessment
- Driver influence - The extent to which the driver influences the FoM.

From the Figure 2.1 it can be seen that speeding has a high relevance on driver's behaviour. Other parameters such as swerving and cornering are considered for scoring the driver's behaviour. Swerving is defined as an abrupt change in direction of the vehicle, for example number of rapid lane changes while overtaking. Cornering is defined as a turn taken at high speed which increases the risk of vehicle roll-over or skidding.

Relative speeding in reference to the legal speed limits of the road is obtained by reverse geo-coding the GPS co-ordinates which is provided by *OpenStreetMap* API. Over-speeding may be detected by comparing the road data with the vehicle speed at a certain location. Percentage of the time in which the driver exceeds the limits is used to determine the magnitude of over-speeding. Smoothness in driving is a measure about the driver's anticipation and the ability to cruise for longer duration. A driver who achieves a higher average speed may not be regarded as aggressive. Whereas the density vs speed distribution can be used to verify the smoothness in the quality of driving.



CHARACTERIZATION OF FOMS IN INSURANCE TELEMATICS CALCULATED USING GNSS-DATA IN TERMS OF FOM OBSERVABILITY, EVENT STATIONARITY, DRIVER INFLUENCE, AND ACTUARIAL RELEVANCE. CHARACTERIZATION PERFORMED AS "LOW", "MEDIUM", OR "HIGH".

FoM	Description	FoM observability	Event stationarity	Driver influence	Actuarial relevance
Acceleration	Number of rapid acceleration events and their harshness	Medium	Low	High	Medium
Braking	Number of harsh braking events and their harshness	Medium	Low	Medium	High
Speeding (absolute)	Amount of absolute speeding	High	High	High	Medium
Speeding (relative)	Amount of speeding relative a location dependent limit	Medium*	High	High	High‡
Smoothness	Long-term speed variations around a nominal speed	High	High	Medium	Low
Swerving	Number of abrupt steering maneuvers and their harshness	Low†	Low	Medium	Low
Cornering	Number of events when turning at too high speed and their harshness	Medium	Medium	High	Medium
Eco-ness	Instantaneous or trip-based energy consumption or carbon footprint	Low	Medium	High	Low
Elapsed time	Time duration of the trip	High	High	Low	Low
Elapsed distance	Distance of the trip	High	High	Low	High
Time of day	Actual time of day when making the trip	High	High	Low	High
Location	Geographical location of the trip	High*	High	Low	Medium

† Not observable using only GNSS-receiver data. Fusion with inertial measurements is required.
‡ Given that the database with speed limits is sufficiently accurate.
* Digital map or database required.

FIGURE 2.1: FoM table for aggressive driving [13]

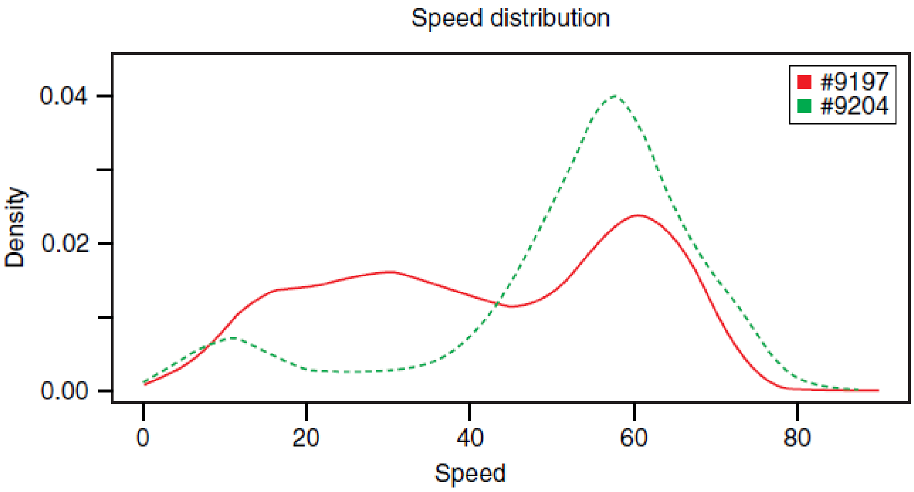


FIGURE 2.2: Density Plot of Speed Distribution [20]

Figure 2.2 shows density vs speed plot comparison of two drivers. It can be seen that #9197 travels at slower speed but accelerates and decelerates frequently. This could also mean that #9197 is experiencing more traffic.

2.2.1.2 Influence of Speed on Eco-driving

Speed has a direct influence on the fuel efficiency. Driving at very high speeds or very low speeds will result in increased fuel consumption [21]. Driving speed and fuel efficiency form a U-shaped curve. There is a certain range at which the fuel efficiency is maximum and it is dependent on the vehicles engine specifications. Steady driving i.e. cruising at a constant speed using inertia contributes to reduced fuel consumption [22]. The speeds in the U-shaped curve are usually below the speed limits of highways. European Environment Agency showed that reduction in speed limits on a highway by 10 km/hr increased the fuel efficiency by 10% for diesel and 18% for gasoline cars [7].

2.2.1.3 Statistical Features

For classification of the driver based on the speed several statistical features are derived from this temporal signal. They are Mean/Average, Standard Deviation, Minimum, Maximum, Kurtosis and Skewness.

Skewness is a measure of symmetry of a data set. Histogram of a perfectly symmetric data will have a skewness of 0. Kurtosis indicates the "peakedness" of the data. A data set can be classified as platy kurtic, normal kurtic or lepto kurtic. Using these classifications of speed data based on the mentioned kurtosis distributions driving can be classified as highway/urban style. A highway driving will have kurtosis greater than 3 (k > 3) and an Urban driving with kurtosis less than 3 (k < 3) will be classified as platy kurtic [23].

2.2.2 Acceleration/Deceleration

Aggressive driving style is characterized by hard acceleration and deceleration. To gain speed or reduce speed one accelerates and decelerates. The aim of eco-driving is to reduce the magnitude of these variations.

External sensors such as accelerometers and gyroscopes can be used to measure the acceleration and differentiating speed acquired from a GPS also results in acceleration/deceleration value. Deceleration is negative acceleration, in transportation terms it is called as braking. Unit of acceleration is m/s^2 or *km/hr/s*. Further, acceleration can be classified into longitudinal and lateral types. Swerving behaviour is captured by a high sample rate accelerometer and gyroscope [24], in this paper the authors have measured the curvature of the curves to classify the maneuvers as swerving, lane changes, and parking. Figure 2.3 shows the plots for the x-axis of gyroscope. It can be seen that the swerving and turning can be differentiated based on the sign changes and the amplitude of the signal.

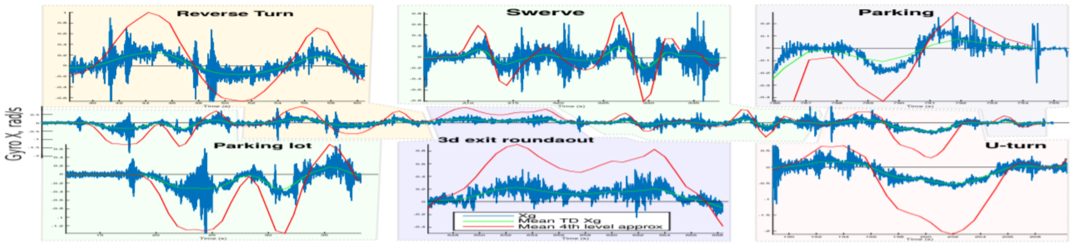


FIGURE 2.3: Various maneuvers captured by accelerometer and gyroscope [24]

Driving features extracted from the sensors support the fact that aggressive drivers have higher g-forces for both acceleration and deceleration [25]. Sudden changes in speed while driving indicates an aggressive driver pattern. In [26] the authors have considered 2.74m/s^2 as the limits for hard acceleration and deceleration, normal acceleration/deceleration from 0.1m/s^2 to 2.74m/s^2 and cruising range from 0.1m/s^2 to -0.1m/s^2. Other papers have used similar limits in defining the aggressive behaviour of the driver. Statistical features extracted from acceleration are average, maximum, standard deviation, skewness and kurtosis.

2.2.2.1 Impact on Eco-driving

Fuel efficiency is maximized when acceleration and braking are reduced. Accelerating to higher speed reduces the time to anticipate the traffic ahead leading to application of brake and hence losing the energy gained to attain a certain speed. Instantaneous fuel consumption is higher during acceleration as opposed to deceleration. During acceleration the engine needs continuous fuel supply, but during deceleration (without brake and accelerator pedal pressed) fuel is only necessary to keep the engine from turning off. This kind of deceleration is referred to as engine braking.

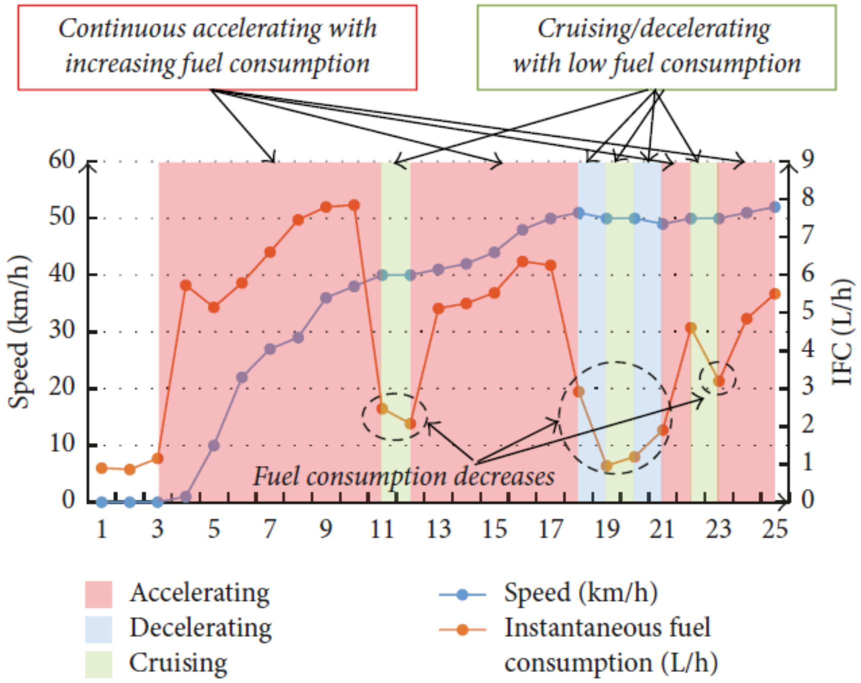


FIGURE 2.4: Fuel Consumption during acceleration, deceleration and cruising [27]

In Figure 2.4 it can be seen that the acceleration has higher instantaneous fuel consumption compared to deceleration and cruising. However, sharp deceleration indicates lack of anticipation and has an effect on the overall fuel consumption due to the loss of inertia [27].

2.2.3 Engine Speed and Fuel Consumption

OBD-II provides engine sensor related information such as -

- Throttle Position
- Manifold Pressure
- Mass Air Flow (MAF)
- Injection Valve
- Idle Speed Valve

Engine Speed is a factor that has a significant weight in fuel consumption, for this reason engine speed is often one of the first parameters that are examined for evaluating fuel efficiency . Fuel consumption is often represented as the amount of fuel (in Litres) consumed per 100 km. A gasoline engine uses a spark plug to trigger the combustion of fuel and air mixture, whereas a diesel engine compresses air and fuel until it reaches a temperature for self-ignition. For a given speed instantaneous fuel consumption is calculated as the ratio of fuel flow to the speed [28]. In many vehicles, fuel rate data is not available due to the missing sensor between fuel tank and engine or the manufacturer chooses not to make the data available. But using MAF, fuel flow is calculated by considering air-to-fuel ratio and fuel density [29].

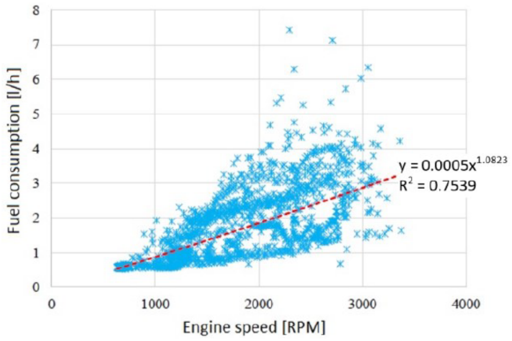


FIGURE 2.5: RPM vs Fuel Consumption [29].

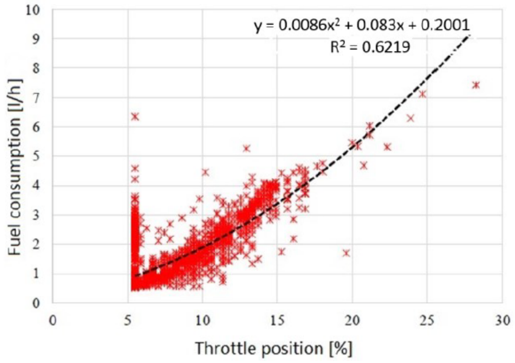


FIGURE 2.6: Throttle Position vs Fuel Consumption [29].

Figures 2.5 and 2.6 show the dependence of fuel consumption on engine speed and throttle position. It can be seen that there exists a linear relationship between these variable. In the Figure 2.6 it can be seen that the fuel consumption at 0% throttle reaches nil value. Automotive manufacturers use engine braking i.e. coasting in a gear with throttle closed to achieve higher fuel efficiency. Engine Control Unit (ECU) takes input from several different sensors such as MAF, Throttle Position Sensor (TPS), manifold pressure, engine RPM, gear, and air intake temperature among several other parameters considered. ECU then takes these inputs and computes the amount of time fuel has to be supplied to the engine to keep it from turning off. So, coasting in neutral consumes more fuel than in engine braking [30].

Idling is defined as the time spent in keeping the engine running with zero distance covered. It is more fuel efficient to switch off the engine than leave the engine running [31]. A car that is idling typically causes 0.4 g/s of CO2 emissions, Reducing idling time, by turning off the engine might therefore lower fuel consumption. In the Table 2.1 taken from [32], shows the amount of time a driver spends in idling state impacting the fuel efficiency, emitting greenhouse gases with zero distance traveled.

Road Type	Average Idling	Min. Idling	Max Idling
Urban	15%	0%	50%
All	10%	0%	37%

TABLE 2.1: Percentage of Idling Time [32]

2.2.3.1 Gear Shifting and Engine Speed

Choice of shifting gears at a certain engine speed is personal and distinguishable among other eco-driving parameters. It is influenced less by the driving context, but can be affected by the power-to-mass ratio of the vehicle. A higher power-to-mass ratio enables the vehicle to be driven at higher gear at low speeds. An eco-driver will shift into higher gear at a certain RPM (usually lower). In [11] the authors have listed four golden rules for eco-driving. Two of the rules are related to gear shifting and RPM. They are -

- **Shift up as soon as possible** - Shifting to a higher gear around 2000-2500 RPM will result in a better fuel efficiency
- **Maintain a steady speed** - Use the highest gear possible and drive with low engine RPM.

Every gear corresponds to a fixed ratio, while gear shifting or braking results in intermediate speed and RPM values. Driving style is characterized by the moment just before changing the gears indicated by velocity and acceleration.

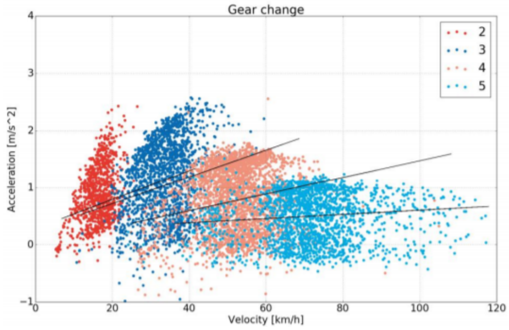


FIGURE 2.7: Velocity vs Acceleration for different gears. Representing good eco-driving behaviour [32].

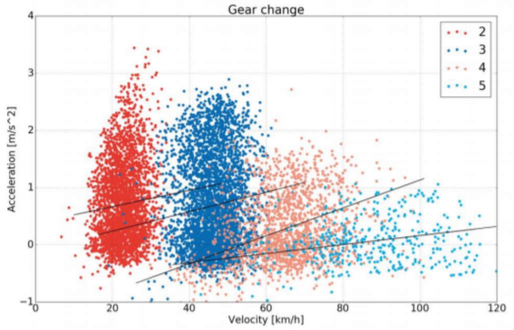


FIGURE 2.8: Velocity vs Acceleration for different gears. Representing bad eco-driving behaviour [32].

In the Figures 2.7 and 2.8 from [32] shows the plots for eco-driving styles related to gear changes. Shifts to different gears are indicated by colours, bad eco-drivers tend to accelerate more and then shift to higher gear. From this study the authors found that there is a large bandwidth in average RPM at which the gear transitions from lower to higher gear among the drivers. This large bandwidth means there is quite some room for improvement by driving efficiently. In terms of fuel consumption, the lower the RPM the better. The average engine speed, acceleration and velocity and the slope of the fit are performance indicators of the gear changing behaviour. These parameters are to be considered as one of the eco-driving parameters.

2.2.4 Positive Kinetic Energy and Relative Positive Acceleration

Positive Kinetic Energy (PKE) is associated with the driver's anticipation of the traffic. It is the ratio of all the positive accelerations encountered to the distance traveled. This variable measures the aggressiveness of driving and depends on the number of variations and magnitude of speed. In [11] the authors have considered PKE as one of the performance indicators for golden rules of eco-driving. PKE represents the driver's ability to keep the kinetic energy as minimum as possible. A rash driving will be associated with a high PKE and a smooth driving will have a low PKE [11] [33].

Relative positive acceleration (RPA) is defined as the product of the instantaneous speed and the instantaneous positive acceleration divided by total trip distance. This variable is commonly used for validating the driving behaviour in emission testing. In emission legislation, the product of velocity and positive acceleration ("vapos") is often used as an indication of how aggressive the driving style was during a trip [32].

$$RPA = \frac{\sum(v * a)}{D} \tag{2.1}$$

In the Equation 2.5 v and a are instantaneous velocity and accelerations respectively. Lower RPA indicates that the driver is less aggressive.

2.2.5 Weather

Weather can have a direct or indirect influence on vehicle fuel consumption. Rain and snow accumulate on the road and this changes the friction between the wheels and tarmac. As a result wheel slippage increases in turn affecting fuel efficiency [34]. In [31] it is reported that the weather affected fuel consumption of public transport buses in two ways, they are -

- On hot days heavy usage of air-conditioning caused a drop in fuel efficiency
- Heavy rainfall caused traffic jams which led to longer idling duration.

In [15] authors have considered a speed reduction factor for different intensities of rainfall, every segment is enriched with weather conditions through calls to Weather API.

2.3 Related Work

The impact of eco-driving advices and training is experimented in [11]. Two experiments were performed in which the first experiment was only to provide the advice and in the second experiment a group of drivers were given training related to eco-driving. The advices that were provided are according to the eco-driving rules which are described in the previous section. The metrics used to measure the eco-driving behaviour are positive kinetic energy, gear shifting and engine braking. The results of the experiments found an average reduction of 12% fuel consumption. The authors used logistic regression to estimate the influence each of the parameter has for the eco-driving behaviour. It was found that average RPM shift up, Index Gear RPM and PKE are most significant. As these parameters contribute in evaluating the driver, they may as well be used for scoring applications in eco-driving. Idling and high RPM driving has not been taken into consideration in this paper, which also has an impact on fuel consumption.

In [35] the authors have considered driving below speed limit and elimination of idling as the advices along with anticipation, cruising and acceleration behaviour for eco-driving. CAN bus and GPS signals are used to evaluate the data set from four vehicles. Radar plot as seen in Figure 2.9 is used to evaluate the eco-driving behaviour of the four vehicles. Evaluating based on the advices provided, Vehicle 3 performs the worst as it idles for long duration, accelerates at higher RPM and uses more fuel while accelerating. Vehicle 2 is very fuel efficient as it cruises for longer duration and ranks higher for most of the parameters except for the speed limits. Vehicle 1 is best at driving within speed limit and accelerating moderately but mostly drives through city areas. Vehicle 4 idles the least but ranks third on all other parameters. From these observations the authors conclude that not just one parameter on it's own but a combination of all of them contribute to achieving a good fuel efficiency. These two related works have used several features that describe the eco-driving behavior.

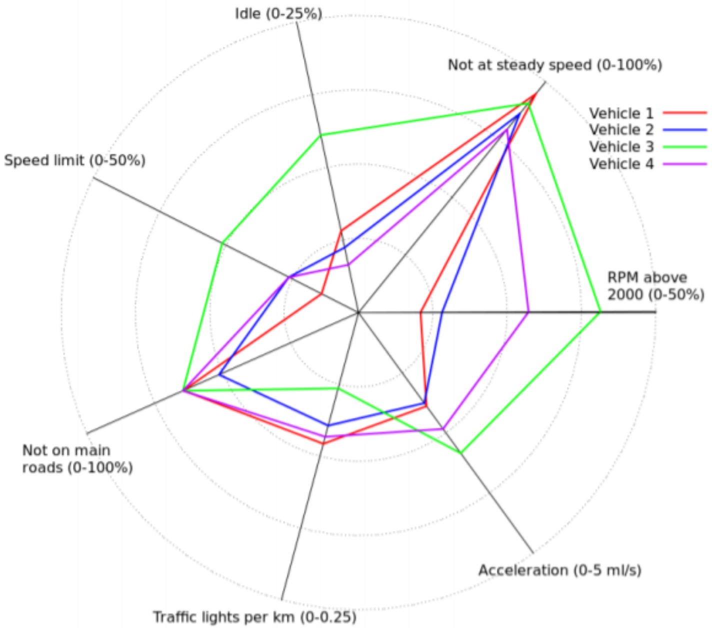


FIGURE 2.9: Evaluation of Eco-driving Advice [35]

Data mining techniques have been used to differentiate and alert the drivers about their driving behaviour by Constantinescu et al. [36]. In-house developed GPS is used to acquire the driving information from 23 drivers. In this paper the authors have used statistical features derived from speed, acceleration, and braking to classify the driving behaviour using hierarchical clustering algorithm. The drivers were categorized into 5 groups of aggressiveness, ranging from very aggressive to non-aggressive. To reduce the number of variables, principal component analysis was used. The authors use principal component analysis to identify the significant features.

UDRIVE [32] is a large scale European project that has collected naturalistic driving data in different countries. In this project the authors have analyzed various eco-driving related factors such as braking, gear shifting and choice of speed on the roads. Based on these factors the drivers and their behaviours are compared country wise. The researchers in this project try to correlate eco-driving and safe driving, but they also mention that there existed no method to measure the safety aspect at that time. Bijman et al. have differentiated safe and unsafe driving using statistical features of speed and acceleration [16]. In this research clustering algorithm is used to classify the driving behaviour. Based on the statistical tests they found that strong accelerating, harsh braking and standard deviation of acceleration as the most important features. As a result, from these features four clusters were obtained. Among these four clusters, one cluster clearly represented safe driving behaviour and one, compared to the others, showed unsafe driving behaviour indications. One of the future recommendations of the author is to use a dimensionality reduction technique to reduce the features and observe if it results in good clusters.

Eco-driving profiling has been combined with Gaming concepts to provide feedback to the driver [37]. In this state-of-the-art research paper the authors have considered fuel efficiency and throttle position sensor as the most important features to profile a driver behaviour. Driving events have been categorized as eco and non-eco driving events. From the naturalistic driving data that has been collected through *enviroCar* database, the authors have proposed a mobile app that alerts the drivers in case of an aggressive event. To give a final score to the driver, throttle position sensor and fuel efficiency is combined. The flowchart of the algorithm is as seen in Figure 2.10. The scoring algorithm represented in this paper gives a final score based only on the fuel efficiency and throttle position values. This enables the driver to compete and get a higher score, but the reason for a particular score is difficult to arrive at just based on the two parameters.

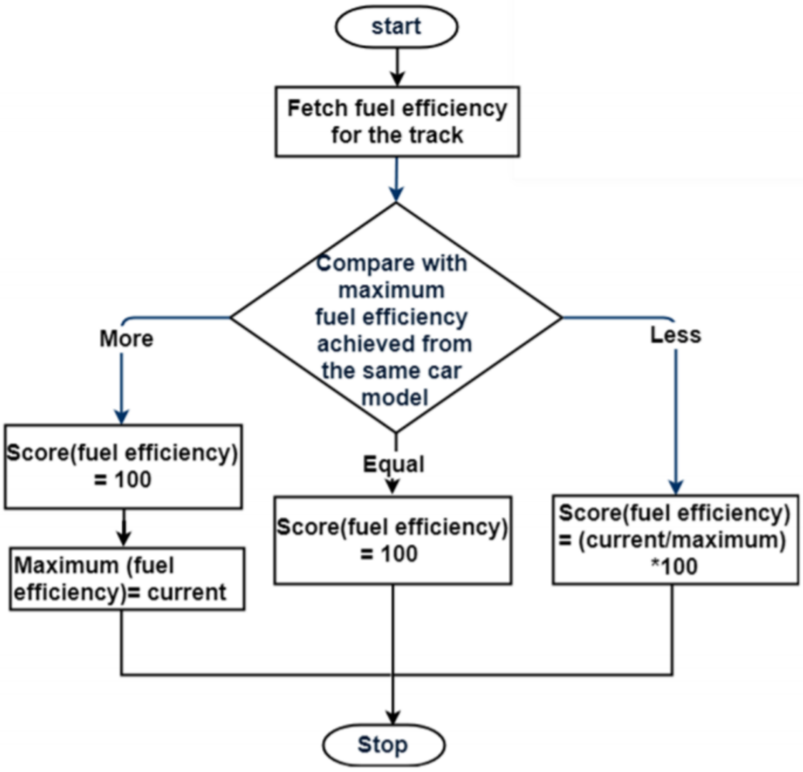

start
Fetch fuel efficiency for the track
Compare with maximum fuel efficiency achieved from the same car model
More
Less
Equal
Score(fuel efficiency) = 100
Maximum (fuel efficiency)= current
Score(fuel efficiency) = 100
Score(fuel efficiency) = (current/maximum) *100
Stop

FIGURE 2.10: Fuel efficiency score algorithm [37]

2.4 Theory

In this section the theory required to understand the methods used in this thesis are presented.

2.4.1 Fuzzy Logic [38]

Fuzzy, the name itself represents that it is imprecise. Fuzzy sets form the base of fuzzy logic. The universe of discourse for classical sets is split into members and non-members. Therefore, classical sets can only represent either '0' or '1', whereas, using fuzzy logic the range from 0 to 1 can represent multiple values. Which means that fuzzy values can be partially true as well as partially false, similar to the analogy of "The glass is half-full" and "The glass if half-empty".

A classical set 'A' is represented in an universe 'U' by membership function

$$\mu_A(x) = \begin{cases} 1, & x \in A \\ 0, & x \notin A \end{cases} \tag{2.2}$$

and for a fuzzy subset 'A' in the universe of discourse 'U' the membership function $\mu_A(x)$ is as shown in equation 2.3

$$\mu_A(x) = \begin{cases} 1, & x \in A \\ 0, & x \notin A \\ (0,1) & x \textit{ possibly belongs to } A \textit{ but not sure} \end{cases} \tag{2.3}$$

This definition allows to map linguistic names to the fuzzy set. These names are adjectives that characterize the fuzzy set.

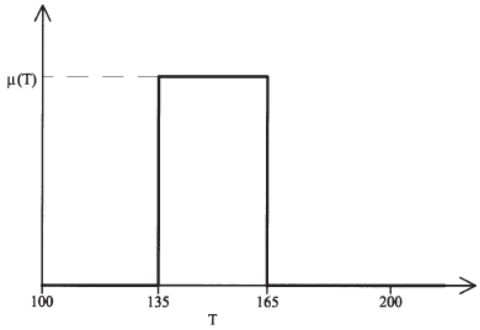


FIGURE 2.11: A representation of classical set [38]

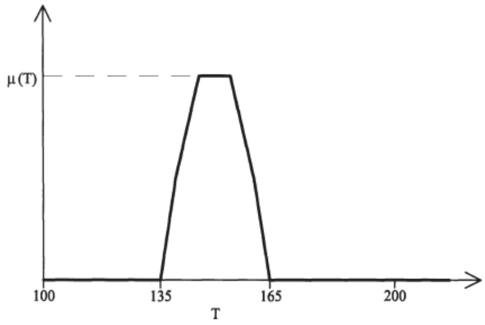


FIGURE 2.12: A representation of fuzzy set [38]

An example of classical set representation and fuzzy set representation is as seen in Figures 2.11 and 2.12 respectively. The boundaries of temperatures in classical set are distinct and precise, whereas in the fuzzy set it is represented vaguely, similar to the way in which a human brain processes a range of variables. In this example, we have seen how to represent certain temperature using a single fuzzy set. Figure 2.13 represents temperature using three fuzzy sets *low*, *medium* and *high*. The shape of the membership function is subjective and it can be applied as per the human decision making capabilities. This is one of the advantages of using fuzzy logic.

Using fuzzy sets requires certain rules to be established. A set of implication forms a rule base. A simple example of an implication can be seen below:

$$If\ x\ \textbf{is}\ A\ \textbf{then}\ y\ \textbf{is}\ B; \tag{2.4}$$

Where A and B are the fuzzy sets. The implication can be divided into two parts. ***If x is A*** will be the antecedent and ***then* y *is* B** is the consequent. Determining the

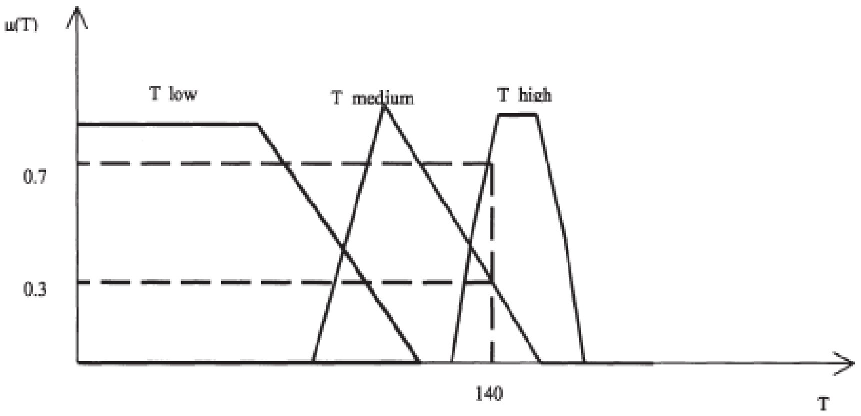


FIGURE 2.13: Representation of multiple fuzzy sets[38]

degree of truth or degree of membership to which an antecedent belongs is called as fuzzification. The values to the degree of membership can be in the range $(0, 1)$. If the value is '0'then it does not belong to the fuzzy set, if it '1'then it completely belongs to the fuzzy set, anywhere in between '0' and '1' represents a degree of uncertainty. In the figure 2.13 a temperature of 140 belongs to *high* with a membership of 0.7 and to *medium* with a membership of 0.3.

Commonly used shapes to define input and output membership functions are:

- Singleton
- Gaussian
- Trapezoidal or triangular

Defuzzification

Defuzzification is the process of obtaining a crisp value from the fuzzified input sets. One of the most commonly used defuzzification method is Centre of Gravity. In this method, input membership function chops the output membership function at the fuzzified values. Then the area under the curve of the output membership function is computed, and the centroid value of this shape is the crisp value, which represents the output.

2.4.2 Supervised Learning

Supervised learning is a type of machine learning algorithm. In supervised learning, algorithms learn to fit a function that maps an input space to an output space. Given a set of input variables 'x', the learning algorithm fits a function to predict the output variable 'y'.

$$y = f(x) \tag{2.5}$$

The data set for supervised learning is split into training and testing data. The training data consists of inputs and respective labels, which are the outputs that are to be predicted on the test data. The main goal of a supervised learning model is to predict the output variable when it is tested on a new set of input data. For this the model has to be generalized. What it means is that, the model should not fit exactly to the training data, which is called as over-fitting. Also, the prediction of the output should not deviate largely from the expectation, called as under-fitting. These concepts can be further explained clearly using bias and variance.

Bias is the difference between the correct value that is to predicted and the average prediction. A large prediction difference between the target and actual value results in high bias. Variance explains the spread of the data on which the model has been trained. When the model fits exactly to the training data, it also fits to the random noise that comes along with it which leads to high variance. An under-fitting model will have high bias and an over-fitting model will have high variance. Bias can be reduced by increasing the complexity of the model, and variance can be reduced by increasing the size of the training data. An optimal model will have low bias and low variance. To minimize the prediction error one has to also consider the bias-variance trade-off.

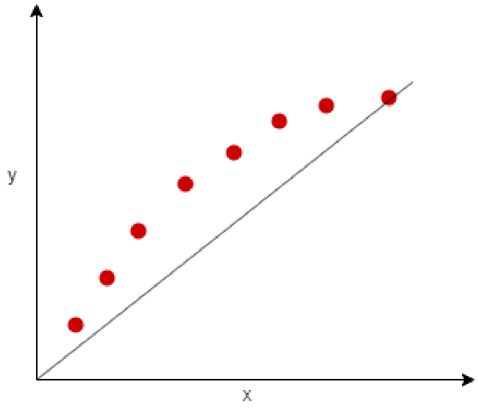


FIGURE 2.14: A model with high bias.

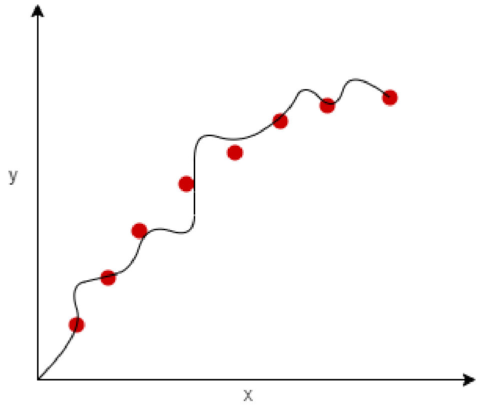


FIGURE 2.15: A model with high variance

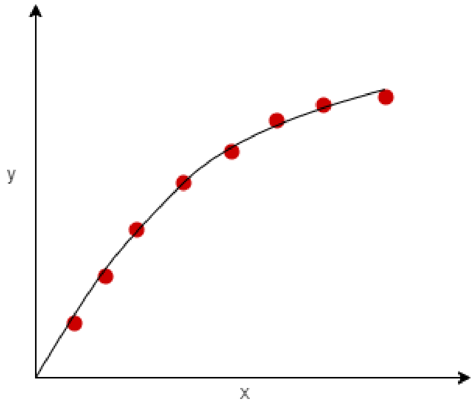


FIGURE 2.16: A model with low bias and low variance

To give an example for supervised learning, let us consider a driver score rating algorithm based on a set of inputs. The range of driver scores are on a scale of 0-10. A driver's profile is described by a number of features such as average speed, acceleration and average RPM. This forms the input space of the model. When the supervised learning model is trained on this type of information, it learns from the observations in the input and output space. Given a set of new data the supervised algorithm will be able to predict the score of the driver.

There are two types of supervised machine learning algorithms:

- Classification - In this type of supervised learning the output space contains labels or classes such as "car", "bike", "truck".
- Regression - In this type of supervised learning the output space has a continuous variable such as fuel consumption.

Decision Trees

Decision trees are a type of supervised learning algorithms used for classification and regressions tasks. They form a tree like structure with a root node, branches and leaf nodes. The nodes in the tree represent the features and the branches represent the values. Data is broken into fragments and with that the tree grows longer. Traversing along the decision tree will lead to a classification.

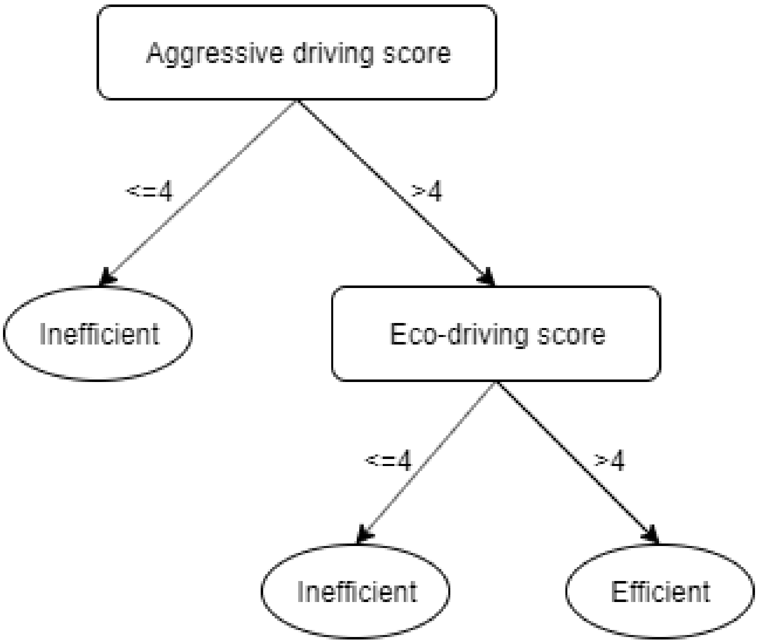


FIGURE 2.17: An example of decision tree

To give an example for the decision tree, let us consider an example of driving style. The score range is from 0-10, where 0 is the least and 10 is the best. In this data set the driver has to classified as *efficient* or *inefficient*. If a decision has to be made, the tree has to be followed from the root node until the leaf. When there is a new observation (instance), if the aggressive driving score is less than 5 then the result is *inefficient*. If the score is greater than or equal to 5 aggressive then the decision is passed to the next node. If the eco-driving score is greater than or equal to 5 then the driver is classified as *efficient*. In this way the decision tree makes predictions about the future instances.

One of the major benefits of the decision trees is that they are easy to interpret. But the decision trees are prone to over-fitting by creating complex trees. This can be avoided by ensemble learning methods which will be explained in the next section.

2.4.2.1 Ensemble Learning

In supervised machine learning, models are combined to achieve better performance. This combination is termed as *Ensemble* [39]. There are different types of Ensemble Learning models categorized based on the way they are combined:

- Voting - Models are combined to achieve better predictions. There exists two types of models. Soft voting is a type where the average predictions of the different models are combined to achieve a better performance. In Hard voting the accuracy of the best model is selected from the ensemble.
- Bagging - In this type of ensemble learning, the data set is split into several subsets. To predict an instance the model averages the probabilities all of the models. This way the variance is reduced significantly.
- Boosting - The models are sequentially arranged to form an ensemble. The model adjusts its weights according to the prediction. If the prediction is correct, it decreases the weight, otherwise it increases the weight. Finally all the models weigh in together to predict an instance.
- Stacking - A stack of supervised learning models also called as base learners are trained on a set of data. The output of this prediction is fed as input to another set of models. These models predict the output, if it is satisfactory. Otherwise the number of stacks may be increased.

In order to increase the accuracy of prediction, ensemble models are used in machine learning often. One of the most commonly used machine learning model is Random Forest. In this thesis, this model is used to evaluate the feature importance.

Random Forest

Random forest is a collection of many decision trees. Decision trees are prone to over-fitting individually [40]. As explained earlier, this leads to high variance and in turn causes errors in prediction. A collection of decision trees will solve the problem of higher variance. Every decision tree is trained on a random subset of data. The errors caused by each decision tree is random and when they are averaged what remains is the actual prediction that was desired. Figure 2.18 represents an ensemble of decision trees.

Feature Importance

An advantage of using ensemble models is that they provide the importance for variables used to predict the output. Feature importance is calculated by the number of times a feature is used to split a node. To give an example for the usage of feature importance, let us consider driving related features as inputs. The features can be acceleration, deceleration, speeding among others. The output variable is fuel consumption. When an ensemble model is fit to this data set, it is possible to see which feature has more predictive power over others. We will use this feature to determine the weights of the features in our scoring algorithm. It will be discussed in the next chapter.

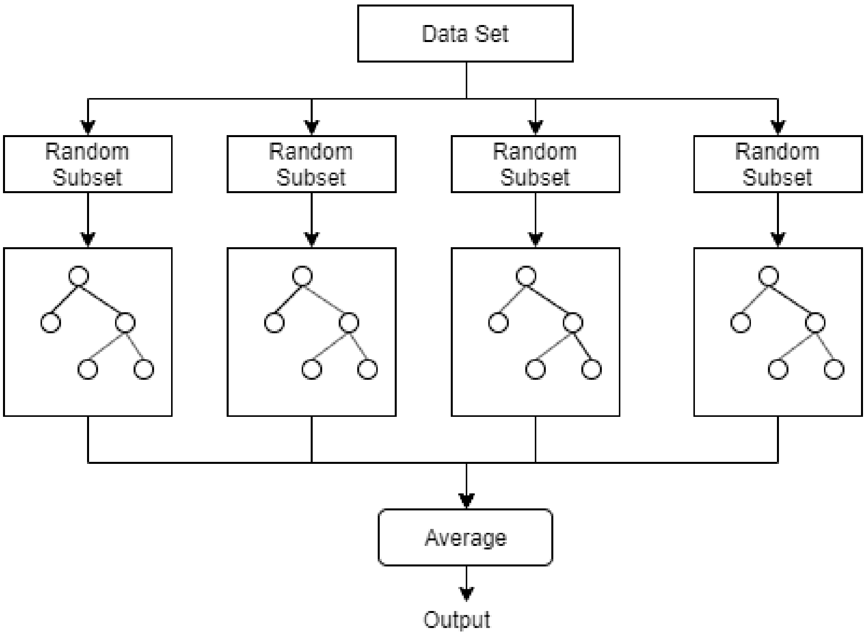


FIGURE 2.18: Random Forest

2.4.3 Unsupervised Learning

Unsupervised learning is a type of machine learning algorithms where the data set comprises of only inputs. Which means that the data set has no labels and classes. In supervised learning we have seen that there exists a ground truth to train the model and later make predictions on the data set. Therefore, the approach taken to perform the learning tasks in unsupervised learning is very much different from that of supervised learning. Algorithms are left to themselves to perform the task and discover the patterns in the structure in the data. One has to take heuristic approach to validate the tasks performed by unsupervised learning algorithms.

Unsupervised learning consists of two major types:

- Clustering - It is the process of finding similar patterns in a data. There are different clustering algorithms, and usage of these algorithms is dependent on the type of data.
- Association - Association allows establishing relations within the database. For example, a group of people who prefer electric cars over regular cars.

2.4.3.1 Clustering

Clustering is a process of gathering or grouping data points. It is done is a way where the points belonging to one cluster are very similar and the points belonging to other clusters are dissimilar. In this section, different clustering algorithms will be discussed. Clustering algorithms are one of the common types in unsupervised learning.

Dissimilarity measures

In cluster analysis, to determine the similarity or associativity between two data points one has to specify a distance metric. This is called as dissimilarity measure. Most commonly used distance measures are:

- Euclidean Distance - It is the measure of the distance between two points in two or three dimensional space.

$$d_{euc}(x,y) = \sqrt{\sum_{i=1}^{n}(x_i + y_i)^2} \tag{2.6}$$

- Manhattan distance - In this method the distance is measured along the axes at right angles.

$$d_{man}(x,y) = \sum_{i=1}^{n} |(x_i - y_i)| \tag{2.7}$$

2.4.3.2 K-Means

One of the most popular and commonly used algorithm in unsupervised learning is k-means [41]. The algorithm initially requires k as the number of clusters the data points have to be grouped into. The algorithm starts by placing k centroids randomly. Then, each data point is assigned to the cluster based on the least euclidean distance between the point and the cluster. Let $X = x_1, x_2, ...x_n$ be the set of data points that needs to be grouped into the cluster set $S = s_1, s_2, ..s_k$. The data points have d dimensions. Then, each data point x assigned to cluster k can be expressed as:

$$argmin_{s_i \epsilon S} \left(\sqrt{\sum_{i=0}^{d} s_i - x_i} \right) \tag{2.8}$$

In the next step the centres are re-positioned as per the mean of the data points in that cluster. Hence the name k-means.

$$s_i = \frac{1}{s_i} \sum_{x_i \epsilon s_i} x_i \tag{2.9}$$

These steps are performed until the clusters reach convergence.

K-means is an extremely fast and less complex algorithm compared to other clustering algorithms and has a time complexity of $\mathcal{O}(\log n)$. The drawbacks also have to be considered while using this algorithm. One has to specify the number of clusters which is not known a priori. For this, it requires evaluation methods which will be discussed later in this chapter. Since the measure of the distance between the points in euclidean distance, it is sensitive to outliers. Apart from the drawbacks, the algorithm is relatively simple to interpret and use.

Hierarchical Clustering

Hierarchical clustering is a type of clustering where similar data points are grouped together in a hierarchical manner [42]. It is composed of two main types, they are Agglomerative and Divisive Clustering. These two types of clustering methods depend on the distance metric and linkage criteria, which has to be determined. Distance metric measures the separation between two points, which means the similar points are closer and the dissimilar points are farther from each other. Linkage criteria determines from where the distance is measured between two clusters. For example, we could consider an average/mean distance between two clusters, in this case we re measuring the distance from the middle of one cluster to another. There are other types of linkage criteria which will be explained in this section.

First let's look into the different types of hierarchical clustering.

- Agglomerative Clustering - In this type of hierarchical clustering, every data point is considered a cluster initially. Depending on the distance metric and linkage criteria, the closet points will merge together to form a cluster. This process is repeated until all the data points are assigned to a to one single cluster. The result is represented as a dendrogram, which has a tree like structure.
- Divisive Clustering - In this type of hierarchical clustering, all the data points are initially considered to be one single cluster. Divisive clustering follows a top-down approach. Instead of merging the clusters, in this method the dissimilar points are split into different clusters. This process is repeated until the entire set of data points are on their own.

Agglomerative clustering is applicable for small clusters, whereas divisive clustering is applicable for very large clusters in general.
There are four types of linkage criteria that can be specified:

Let i and j be the clusters and x is the set of data points.

- Single Linkage - In this type of linkage, two points that have the least distance among the clusters are considered as the reference point of measurement. Single linkage is also called as Nearest Neighbour Clustering. For single linkage criteria, the distance function can be expressed as:

$$d_{SL}(i,j) = \min_{x_i \epsilon i, x_j \epsilon j} d(x_i, x_j) \tag{2.10}$$

- Complete Linkage - Here the distance between two points among the clusters are farthest from each other. In other words these points are least similar.

$$d_{CL}(i,j) = \max_{x_i \epsilon i, x_j \epsilon j} d(x_i, x_j) \tag{2.11}$$

- Average Linkage- In this type of linkage the distance is measured from the mean of the clusters.

$$d_{AL}(i,j) = \frac{1}{N_i N_j} \sum_{x_i \epsilon i} \sum_{x_j \epsilon j} d(x_i, x_j) \tag{2.12}$$

- Ward's Linkage - It is the most commonly used linkage criteria which is suitable for various types of data sets. The aim of this method is to minimize intra-cluster variance.

$$d_W(i,j) = ||x_i - x_j||^2 \tag{2.13}$$

DBSCAN

Density-based spatial clustering of applications with noise (DBSCAN), is a clustering algorithm based on the density of data points [43]. In this type of clustering method, the data points that are closely packed are grouped into a cluster. In certain scenarios where the data is non-linearly spaced DBSCAN outperforms centroid and hierarchical clustering methods. For example, if the data set forms concentric circles which are densely packed, then the DBSCAN can clearly cluster inner and outer ring. Whereas other algorithms fail because they require the data to be spread out in spectral manner.

DBSCAN requires two parameters to be specified, they are *eps* and *minPts*. *eps* defines the minimum distance for which a data point will be included in the cluster, and *minPts* represents the number of points in that group. The clusters expand if there are neighbouring *minPts* within *eps* radius. The expansion stops if the criteria is not satisfied. Every point in the data is checked for this criteria. Points that are not in the cluster are considered as noise. DBSCAN is a very powerful algorithm if used with right parameters. One needs to have the knowledge of the dataset to use this algorithm.

2.4.3.3 Cluster Evaluation Techniques

Evaluation of the clusters is important to determine the quality that they represent. If the objects in the cluster are similar, then the cluster quality is high. Low inter-cluster similarities and high intra-cluster similarities is a good representation of the clusters formed in the entire data set.

Silhouette score

Silhouette score is an evaluation technique in which a data point's distance is measured with respect to other data points in the same cluster, and with the data points in the other neighbouring clusters [44]. Quality of the cluster is determined by the score. A higher score indicates a good cluster representation.

Mathematically it can be represented as seen in the Equation 2.14. Let *i* be the data point which belongs to the cluster *a*, and let *a(i)* represents the mean distance between *i* and other points in the same cluster, *b(i)* represents the mean distance between *i* and the neighbouring clusters. Then silhouette score *s(i)* is expressed as -

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))} \tag{2.14}$$

It is clear from the equation that the range of s(i) is [-1,1], where '1' represents the best score and '-1' represents the least score. By measuring the silhouette score for each data point and then averaging it among the cluster population will represent the score for a cluster.

Elbow method

Clustering algorithms such as k-means requires the number of clusters to be specified [45]. This parameter is often referred to as hyperparameter. Elbow method is one of the most commonly used technique to determine the number of clusters.

The metric that determines the elbow point is Within-Cluster Sum of Squares (WCSS). It is the measure of variability of the data points within the cluster. A cluster that is compact will have smaller sum of squares, and a cluster that is spread will have larger value. But the sum of squares increases with the number of data points. Elbow point is located where the WCSS reduces significantly from the previous observation. From Figure 2.19, it can be seen that the WCSS value drops when the number of clusters is 3. It can also be seen that WCSS drops significantly when cluster number is 2. This makes it quite ambiguous to use the elbow method. One has to resort to empirical methods and/or silhouette score technique to determine the number of clusters.

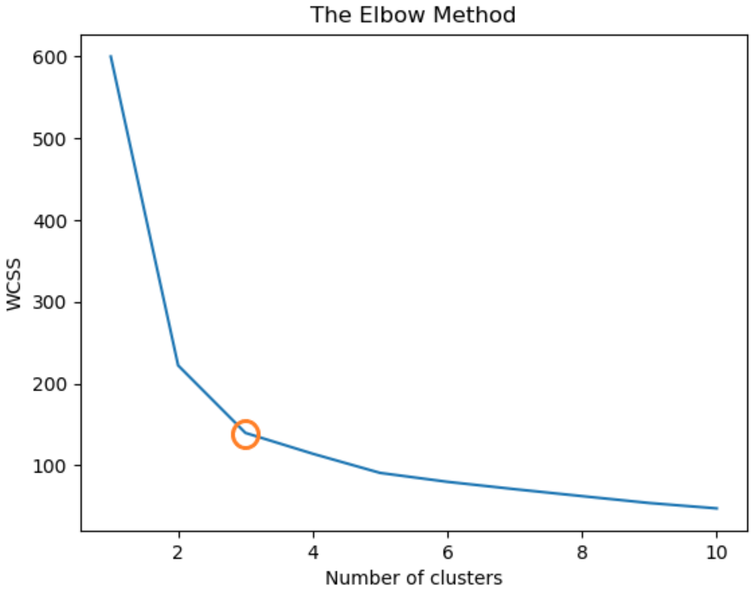


FIGURE 2.19: An example showing elbow point.

2.4.4 Dimensionality Reduction

As the number of features in the data set increases, it becomes difficult to interpret or visualize in 2 or 3 dimensions. Dimensionality reduction techniques reduce the number of features in a data set into principal components such that these components account for most of the information. The number of principal components is much lesser than the original original number of features.

2.4.4.1 Principal Component Analysis

Principal Component Analysis (PCA) is dimensionality reduction technique where a set of possibly correlated variables are converted into a set of linearly uncorrelated variables called principal components [46]. It is a statistical procedure which uses orthogonal transformation, which means each of the principal components are orthogonal to the preceding ones. The first component accounts for the highest variance. The remaining components amount to the remaining variability. The first step is to form a covariance matrix C represented mathematically as:

$$C = \sum_{i=1}^{n} \frac{(x_i - \hat{x}) - (x_i - \hat{x})^T}{n-1} \tag{2.15}$$

where $\hat{x}$ is the sample mean. Next step is to calculate the eigen values (characteristic roots) and eigen vectors (principal components) from the covariance matrix. We obtain eigen values λ by solving the equation below.

$$|C - \lambda I| = 0 \tag{2.16}$$

To obtain eigen vectors v we solve the Equation 2.17.

$$(C - \lambda I)v = 0 \tag{2.17}$$

Every eigen vector has an eigen value, they form pairs. Eigen vector represents the axis with most variance and eigen value is the corresponding coefficient, which represents the magnitude of the variance. Sorting the eigen vectors in the order of their eigen values from highest to lowest, we obtain the principal components.

2.4.4.2 T-Stochastic Neighbour Embedding

T-Stochastic Neighbour Embedding (t-SNE) is a popular dimensionality technique which is applicable on non-linear data sets as well [47]. t-SNE retains the local structure better than other dimensionality reduction techniques. The high-dimensional points create a probability distribution that direct the relationship between other points in the high dimensional space. Then the same probability distribution is followed in the low dimensional space as best as possible to map the neighbouring points.

The conditional probability that a point x_i picks x_j as a neighbour in high dimensional space can be expressed as:

$$p_{ij} = \frac{exp(-||x_i - x_j||^2/2\sigma_i^2)}{\sum_{k \neq i}(exp||x_i - x_k||^2/2\sigma_i^2)} \tag{2.18}$$

The probability function follows Gaussian distribution, therefore σ_i represents the variance of the Gaussian centred on x_i. To capture the local structure well Gaussian distribution is used. In other words, the points that are spaced farther have low σ_i than the ones that are close by. This how t-SNE retains the local structure.

We want to capture the same probability distribution in the lower dimensional space as well. Let q_{ij} represent the conditional probability in low dimensional space, y_i and y_j are the low dimensional counterparts of x_i and x_j. We want q_{ij} to be as similar as possible to p_{ij}. Therefore the conditional probability in lower dimensional space is -

$$q_{ij} = \frac{exp(||y_i - y_j||)^2}{\sum_{k \neq i} exp(-||y_i - y_k||^2)} \tag{2.19}$$

If the mapping between q_{ij} and p_{ij} is modeled correctly then these two conditional probabilities will be equal. t-SNE aims to minimize the difference between them by using a cost function. The cost function used minimizes the sum of Kullback-Liebler divergence over the entire set of data points using gradient descent method. The cost function is mathematically represented as:

$$C = \sum_i KL(P_i||Q_i) = \sum_i \sum_j p_{ij} log \frac{p_{ij}}{q_{ij}} \tag{2.20}$$

Thus minimizing the KL between high dimensional point x_i and low dimensional point y_i using gradient descent results in better mapping.

2.5 Summary

In this chapter, eco-driving and the golden rules of eco-driving were introduced. Followed by that, the parameters that are related with the driving behaviour are discussed. Also, related work on which this thesis is based on was presented. At last, theory of the concepts used in this thesis were presented.

Chapter 3

Methodology

3.1 Overview

Vehicle's fuel consumption and eco-driving is greatly influenced by the driving behaviour. In Chapter 2 eco-driving rules and their implications on the fuel consumption were discussed. These rules are also explored in this chapter. Apart from these rules, this chapter also focuses on driving behaviour classification. This classification is a two step process: step 1 is feature extraction, and step 2 is driving behaviour categorization using scoring and clustering. Feature extraction from the continuous time-series OBD-II data is considered as a challenge. Some features such as Gear Shifting, Engine Braking, Cruising etc. are not readily available from the time-series data and extensive amount of processing is required to extract these features. Although feature extraction is considered as a challenge, it improves the accuracy of the classification model.

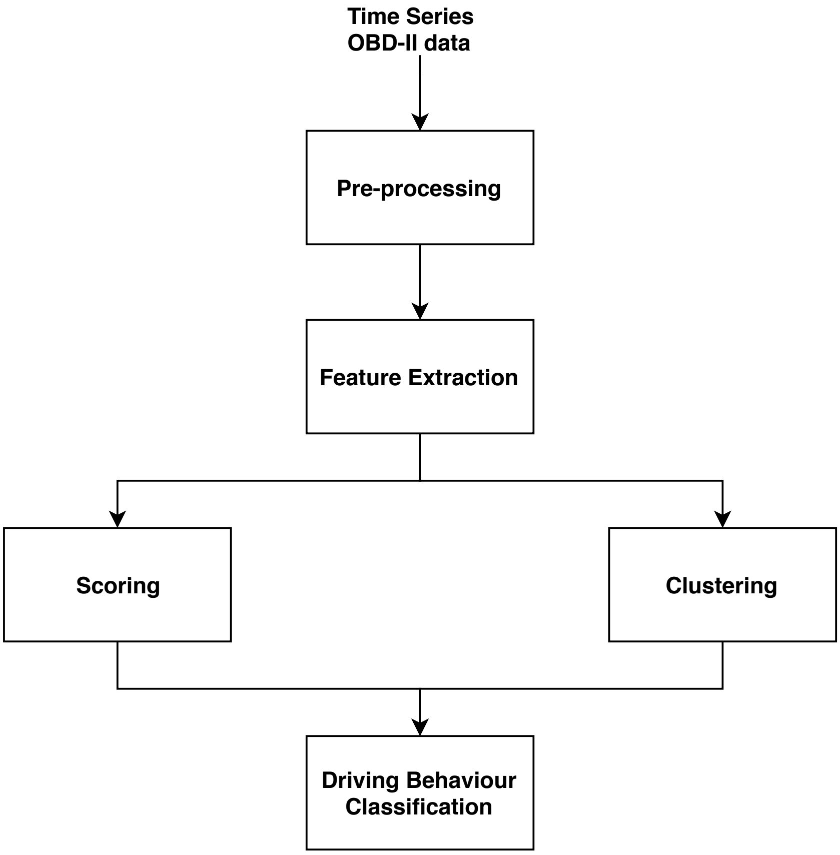


FIGURE 3.1: Brief overview of the methodology to classify driving behaviour.

There are multiple ways to do behaviour classification, for e.g. supervised learning where the ground-truth about the driving behaviour is known. The complexity of driving behaviour makes it difficult to gather the ground-truth labels, making supervised learning a difficult approach for classification. Instead, we use a scoring technique and clustering algorithms (unsupervised learning) to categorize the driver behaviour. Figure 3.1 presents an overview of the work flow used to classify the driving behaviour.

This chapter starts with a discussion about the feature extraction from continuous time-series data. These features along with the fuzzy logic model are used to generate a trip-level scoring algorithm. Later, in this chapter the workflow for clustering algorithm is presented.

3.2 Pre-processing and Feature Extraction

The OBD-II time-series data consists of raw information provided by the vehicle's ECU like speed, engine RPM etc. This raw information is used along with the eco-driving rules to extract the features that characterize the driver's behaviour and affect the fuel efficiency. In this section, these features along with the methods and calculations used to extract them are discussed.

There are a lot of signals that are provided by the vehicle's ECU, but only a few of these signals are utilized here. These signals are listed below:

- **Speed (v)** - The velocity of the vehicle expressed in kilometres per hour (*km/hr*).
- **Engine Speed (RPM)** - Engine speed expressed as the number of revolutions per minute.
- **Throttle Position** - The value of the accelerator position expressed in percentage. The value ranges from 14% (closed throttle) to 100% (fully open throttle). It regulates the rate of air flow into the engine.
- **Mass Air Flow (MAF)** - To maintain an optimal air-to-fuel ratio, the engine uses MAF sensor to regulate the fuel flow. This fuel flow is expressed in grams per second (*g/s*).
- **Instantaneous Fuel Consumption** - Some car manufacturers provide the fuel consumed by the vehicle at a given instant. Fuel efficiency is expressed in Litres per Hundred kilometres (*L/100km*).

The frame rate of the time-series OBD-II signals is not uniform across different vehicles. Depending upon the number of requests to ECU, this frame rate can be as high as 100 Hz (only one signal acquired). At this high frame rate, no significant changes are observed in the signals. Therefore, a frame rate of 1 Hz is chosen which is sufficient to acquire all the signals and observe desired signal changes. The pre-processing stage presented in Figure 3.1 takes care of down-sampling the frame rate to 1 Hz across all the data sets.

3.2.1 Harsh Acceleration and Deceleration

Harsh acceleration and deceleration, as explained earlier in Chapter 2, are some of the important features that are derived from the speed signal v as presented in the

Equation 3.1. These features are indicative of fuel efficiency and aggressive driving behaviour.

$$a = \frac{\Delta v}{t} \tag{3.1}$$

The acceleration/deceleration are classified as harsh based on a minimum velocity change of 11 km/hr per second. This minimum value of velocity change is based on the literature from [26]. To utilize these features in the scoring algorithm, the number of harsh acceleration and deceleration are counted. Since the trips driven can be of varying distances, the counted values are measured as number of events per 100 km. Necessary signal conditioning is done to ensure that an event is not counted multiple times.

3.2.2 Positive Kinetic Energy

In the previous section, harsh acceleration and deceleration were discussed. Apart from counting the number of these events, it is also important to analyze their intensities. The feature discussed in this section takes care of this. It accounts for the intensity of the speed gains. It is generally defined as the energy consumed per unit distance traveled and is abbreviated as PKE. Aggressive drivers tend to accelerate to higher speeds at a faster rate, therefore this feature also represents the aggressiveness for the driving style. Equation 3.2 provides a mathematical representation of positive kinetic energy, where v_f is the final velocity, and v_i is the initial velocity. The denominator D is the total distance traveled.

$$PKE = \frac{\sum(v_f^2 - v_i^2)}{D} \qquad when \quad \frac{dv}{dt} > 0 \tag{3.2}$$

Similar to PKE, another feature viz. NKE (Negative Kinetic Energy) is also present which represents the intensity of harsh deceleration. For the collected data sets, both the energy parameters were calculated and plotted. From the plots it was observed that both the features are highly correlated which indicates that a driving style has similar acceleration and deceleration levels. Therefore, negative kinetic energy is not taken into consideration for determining the trip characteristics. This will also help in reducing the number of features to extract, thus reducing the complexity of scoring model.

3.2.3 High RPM and Idling

High RPM (engine speed) is a parameter that has a huge impact on the fuel consumption [29]. According to the eco-driving rules, the driver is always advised to maintain lower RPM. To classify the engine speed as high or low, there is a threshold value which is dependent on the type of fuel used by the vehicle. Gasoline engines operate at a higher engine speeds compared to the diesel engines. The threshold for different engines is presented in Table 3.1.

Fuel Type	**RPM-Threshold**
Gasoline	3000
Diesel	2500

TABLE 3.1: Threshold values for High Engine Speed

When the engine speed for a given fuel type exceeds the set threshold, it is considered as a high RPM event. Along with the event identification, the time for which the engine is running at higher RPM is measured. This feature is expressed as the percentage of total engine run-time.

Idling, as discussed earlier in Chapter 2, is an anticipatory behaviour of the driver. It is represented as the amount of time when the vehicle speed is zero with a running engine. This behavior leads to unnecessary fuel consumption and can be reduced by turning off the vehicle during long stops. Mathematical representation of this feature is presented below.

$$Event = \begin{cases} v = 0 \ \& \ RPM > 0 : Idle \\ v > 0 : Non - idle \end{cases} \tag{3.3}$$

In the given representation, v is the speed of the vehicle at a given instant. Similar to high RPM event, idling is also represented as the percentage of engine run-time.

3.2.4 Cruise

Cruising is defined as the ability to maintain a constant speed (minimum speed deviation from a center value) over certain duration of time. According to one of the eco-driving rules, maintaining a stable speed is required to achieve a greater fuel efficiency. Cruising is not available as a direct parameter from the time series OBD-II data, but has to be derived from the speed of the vehicle. There are multiple ways in which cruising behavior can be identified. One way is to check the acceleration and deceleration, and if they are within a certain range, cruising behavior is identified. With this technique, it is possible that every short segment will be classified under cruising. But it should not be the case and only the actual cruising should be identified. Therefore, in this section, another approach is mentioned which makes use of the sliding window mechanism.

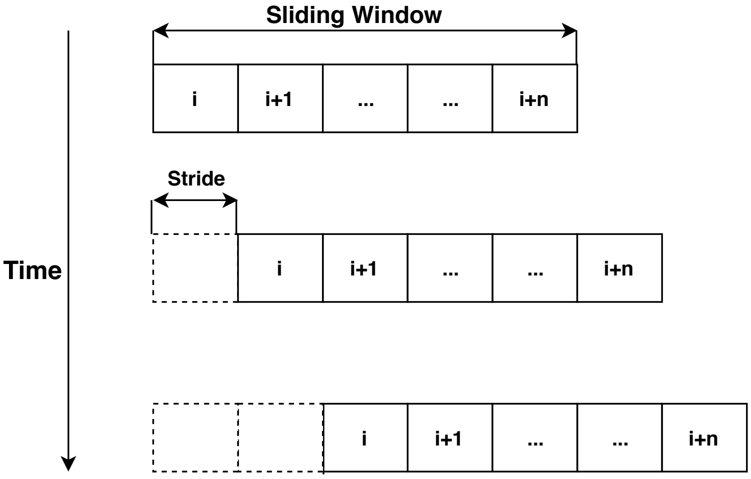


FIGURE 3.2: Sliding window method.

Figure 3.2 presents a generic sliding window of frame size n. At a given instance n values of speed are loaded onto the sliding window and after performing required calculations (explained later), the window slides to capture new data. It is difficult to decide the ideal window size of n but it will be safe to say that if a very short window is taken, very high cruise percentages will be obtained. This might not depict the true scenarios of cruising on a highway. In contrast to this, if a very large window size is taken, then smaller cruise events might not get registered. To tackle with this window size problem, an appropriate window size of 5 is chosen based on some preliminary experimentation.

To classify the sequence of speed values as a cruising event, it has to satisfy certain criteria. As per this criteria, the standard deviation of the speed values in the window should be less than 1 km/hr to register a cruising event. The value for standard deviation is chosen from the literature [27] and verified based on the preliminary experimentation. This sliding window approach is applied at every instance of the time series signal and a counter is used to keep a track of the cruise events.

3.2.5 Engine Braking

Engine braking is generally associated to the anticipatory driving behaviour. Engine braking is an event when the vehicle slows down using the engine's retarding forces instead of the actual braking system. Engine braking differs significantly with respect to the type of the fuel used. In gasoline engines, fuel is cut-off when the throttle position is closed, whereas, in diesel engines there is no concept of throttle valve, so it uses exhaust valve to give the same effect of engine braking. Regardless of the fuel type, engine braking helps in achieving higher fuel efficiency and it also shows the anticipatory driving behaviour. To measure engine braking, two methods are proposed.

- **Method 1** - In this method, instantaneous fuel consumption, mass air flow and speed of the vehicle are considered to calculate engine braking. The main idea here is that the air to fuel ratio is constant during the normal operation of the vehicle, but during engine braking, the fuel flow is cut-off and the air to fuel ratio (AFR) drops significantly. Equation 3.4 shows the formula to compute AFR.

  $$AFR = \frac{Inst.Fuel\ Consumption * Speed}{MAF} \tag{3.4}$$

  AFR is obtained from the standard instantaneous fuel consumption calculations which will be discussed later in the upcoming section. This method is applicable only for gasoline engines as they have fuel cut-off mechanism. In diesel engines there is no intake valve, so fuel cut-off to the engine chamber results in engine braking. The effect of engine braking is higher in gasoline engines compared to the diesel engines due to the fact that fuel-cutoff in gasoline engines cause vacuum development in the engine chamber. In such cases, the engine has to work against the vacuum which causes engine braking. To compute the amount of engine braking, a drop in AFR is considered and all such events are aggregated to compute engine braking for the whole trip.

- **Method 2** - The method discussed earlier works only for gasoline engines where the instantaneous fuel consumption is available. For the diesel engines, a sliding window approximation method is proposed. This approach is similar to the cruising approximation presented earlier. The main differences in both method 1 and method 2 lie in the criterion required to consider an event as engine braking event. In this method, the engine RPM values present in the sliding window should be in the decreasing order to qualify as an engine braking event. Similar behavior is observed when normal brakes are applied. The classification between normal and engine braking is performed by taking into account the standard deviation of the sliding window. The standard deviation for normal braking is significantly higher than engine braking. The sliding window size and the standard deviation values are set by validating against the first method.

Using the above explained methods, engine braking is calculated for both gasoline and diesel engines and is expressed as a percentage of the trip duration.

3.2.6 Gear Shift

Fuel efficiency improves greatly when the gears are shifted at an optimal engine speed. One of the eco-driving rules suggests that the gears should be shifted up at around 2000-2500 RPM. The fuel type of the vehicle also plays a role in determining the shift point. Table 3.2 shows the engine speed limits considered for gear up shifts [48]. The raw time series OBD-II data does not contain any information about the gear position of the vehicle, so, an algorithm is presented in this section to detect gear shifts.

Fuel Type	RPM-Threshold
Gasoline	2500
Diesel	2000

TABLE 3.2: Eco-driving shift up RPM

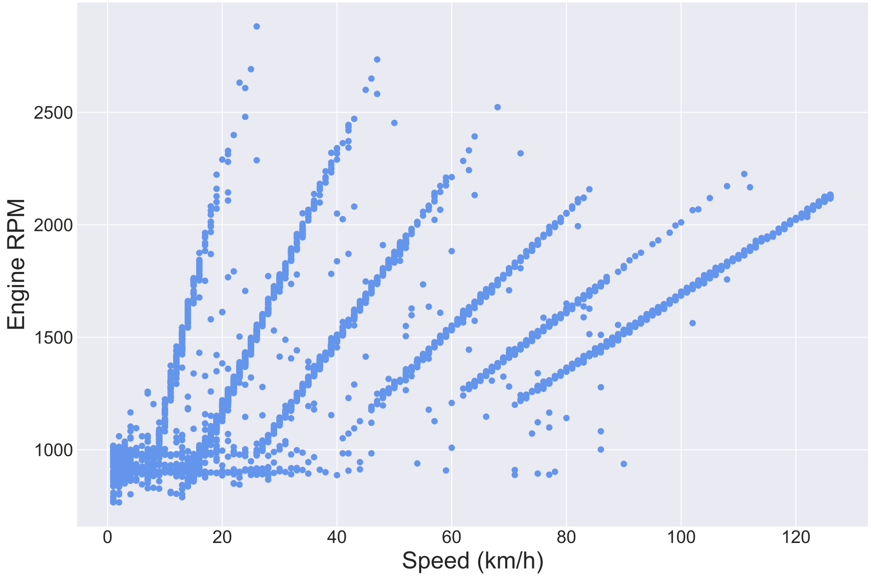


FIGURE 3.3: Speed vs RPM plot.

To understand this algorithm, it is first important to understand what gear ratio means. Gear ratio is defined as the number of turns which input shaft makes to rotate the output shaft once. The speed of the vehicle depends on the engine RPM and the gear ratio which is dictated by the gear under use. To understand more about this gear ratio, a random data set was analyzed and a plot between engine RPM and vehicle speed were plotted. This plot is presented in Figure 3.3. From this plot, it can be observed that multiple straight lines are obtained with unique slope values. Each slope can be identified as a gear present in a vehicle. From this observation, an algorithm was designed to obtain the gear ratios, which is later used to identify the gear.

The steps taken to detect the gear ratios are listed below.

1. Generate an array of RPM to speed ratio.
2. Apply histogram analysis on the array with a bin size of one.
3. Detect the position of peaks in the histogram.
4. Filter the closely spaced peaks, to make the gear ratio distinct.
5. Sort the position of the peaks in descending order to obtain the gear ratio and store it as an array.

The peak positions obtained from the aforementioned algorithm represent the gear ratios of a vehicle. Next task in gear detection is to detect the actual gear in which vehicle is currently driven. This gear position will later be used to detect the transition or switch from lower gear to a higher gear. Figure 3.4 shows the result of the gears identified after applying the gear detection algorithm.

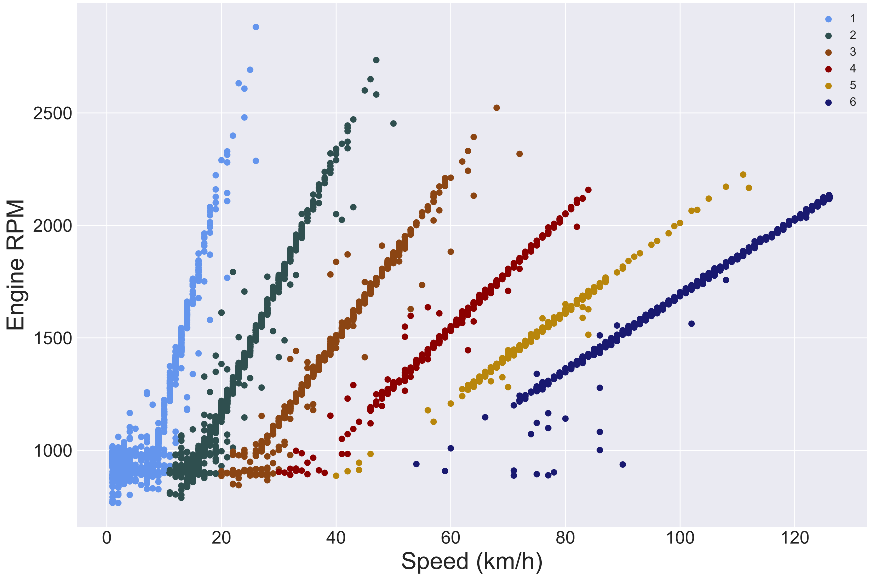


FIGURE 3.4: Speed vs RPM plot after applying the gear detection algorithm

From the detected gears, it is important to detect number of up-shifts present in a trip to evaluate the driver. The process of detecting an up-shift is described below.

1. Compare the current RPM speed ratio to the previously obtained gear ratio array.

2. Assign the respective gear number by finding the nearest value in gear ratio array.
3. When a shift is detected from lower to higher gear, the maximum engine speed around the transition is noted.

From the aforementioned method, engine speed at which up-shifts are performed can be detected. By computing the total number of gear up shifts in a trip, the average RPM at which the gear is shifted and the total number of high RPM gear shifts are derived. The methods described above are applicable to manual and automatic transmission vehicles.

3.2.7 Fuel Consumption

As the name suggests, this feature is very important in detecting the fuel efficiency and classification of drivers. Fuel consumption is calculated using the MAF values. At any given instant when the vehicle is moving the fuel consumed in litres per kilometer is given by the Equation 3.5 [28].

$$Inst.Fuel\ Consumption[l/km] = \frac{Fuel\ Flow[l/hr]}{Speed[km/hr]} \tag{3.5}$$

The above method is applicable only when the vehicle is moving. In scenarios where the vehicle is stationary, fuel consumption is calculated using Equation 3.6.

$$Fuel\ Flow[l/h] = \frac{MAF * 3600}{AFR * FD} \tag{3.6}$$

To perform the above calculations, values for Air-fuel Ratio (AFR) and Fuel Density (FD) for Gasoline and Diesel engines are shown in the Table 3.3.

Fuel Type	**AFR**	**FD** (g/dm^3)
Gasoline	14.7:1	820
Diesel	14.5:1	720

TABLE 3.3: Values for AFR and FD [28]

3.3 Design of Scoring Model

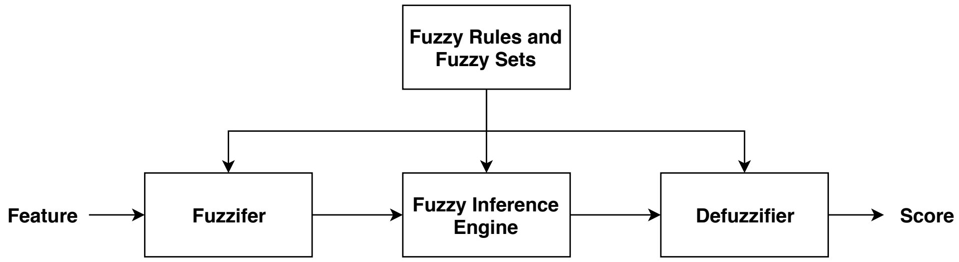


FIGURE 3.5: Fuzzy Inference System

In this section, a fuzzy logic based scoring algorithm is proposed. The trip-level features mentioned earlier are first fuzzified (converting crisp values to fuzzy values) using the defined fuzzy rules. These fuzzified values are then provided to the

Fuzzy Inference Engine. This engine make use of fuzzy values and rule base to perform aggregation. Aggregation is the process by which the fuzzy sets that represent the outputs of each rule are combined into a single fuzzy set. After this step, defuzzification is applied to get an output score. Finally, the trip level scores for all the features are weighted individually and averaged to obtain a single trip-level score. Fuzzy sets, fuzzy rules, and the defuzzification methods will be explained in this section. Later, an architecture for the scoring algorithm will also be presented. An overview of a Fuzzy Inference System (FIS) is presented in Figure 3.5.

3.3.1 Fuzzy Sets

In the scoring algorithm, the trip-level features which were explained in the previous section are used as inputs to the FIS. An input fuzzy set is represented by four linguistic variables: Very Low, Low, Medium and High (*VL, L, M, H*). These linguistic variables enables us to define fuzzy rules, which is later linked to the output fuzzy set. The shape of the input membership function is chosen to be trapezoidal. The limits for each fuzzy set have been defined based on the available data sets. For example, if the maximum idling for any trip in the data set is 20% of the engine runtime then the value is split into four equal parts. This limit (20%) is assumed as the highest possible value that can be attained. For other features such as harsh acceleration, the maximum values are very high for short trips. In this case, the values of all the trips are plotted on a histogram and the peak of the histogram is considered to be the maximum. Figure 3.6 represents a generic input fuzzy set for all the input features. The concept of the membership functions will be explained in the upcoming section.

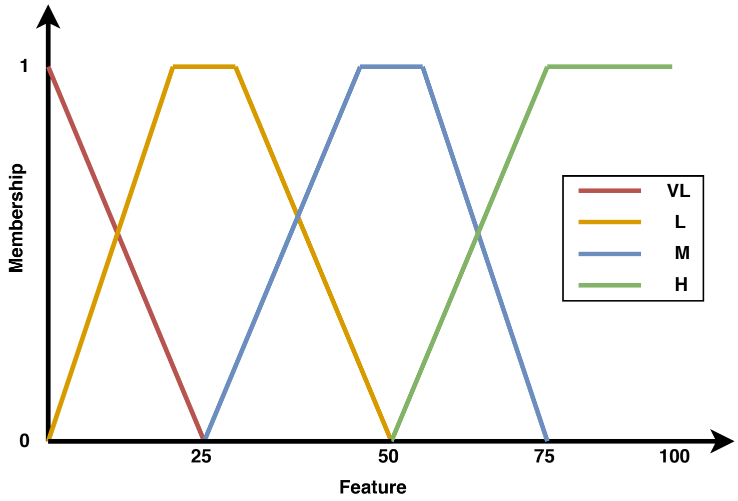


FIGURE 3.6: Input Fuzzy Set

Regarding the output fuzzy set, the score (0-100) is divided into four linguistic variables: Very Inefficient, Inefficient, Efficient, Very Efficient (*VI, I, E, VE*). Figure 3.7 represents the output fuzzy set.

3.3.2 Fuzzy Rules and Defuzzification

The fuzzy rules are defined based on the impact of the feature on the fuel consumption. Basically the input fuzzy set is mapped to the output fuzzy set. For example, if $IDLE = VERY\ LOW => \ SCORE = VERY\ EFFICIENT$. All of the input features

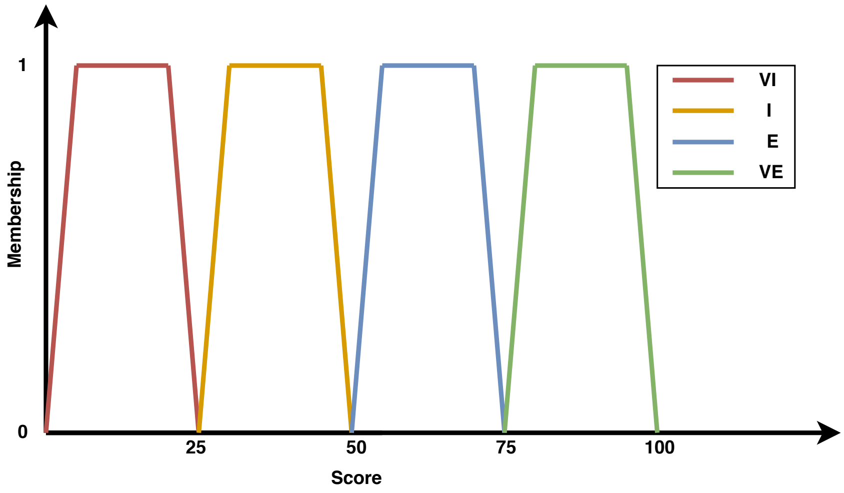


FIGURE 3.7: Output Fuzzy Set

Feature Value	**Score** - Positive Impact on Fuel Efficiency	**Score** - Negative Impact on Fuel Efficiency
High	Very Efficient	Very Inefficient
Medium	Efficient	Inefficient
Low	Inefficient	Efficient
Very Low	Very Inefficient	Very Efficient

TABLE 3.4: Fuzzy rules mapping input membership function to output membership function.

except the features related to gear shifts are single-input, single-output fuzzy sets and require four rules in total. The rules for the gear shift are combined together to give a single score. The number of rules for this case increase from four to eight. An input membership function may trigger one or more output membership function depending on the rules and membership degree. To get a score from a fuzzified value, defuzzification process is applied. The chosen method for the defuzzification is Centre-of-Gravity (COG). This method computes the centroid of the area of the output fuzzy set at which it is triggered (Degree of Membership). If more than one output fuzzy set is triggered then the area covering both the fuzzy sets is taken into consideration. Table 3.4 shows the way the input and output fuzzy sets are mapped to give a score.

3.3.3 Scoring

Now that we have described the methods to provide score for every feature at the trip level, next step is to aggregate the scores to provide the score for the entire trip. A simple averaging is not enough as each feature contributes differently. Some features may contribute heavily in determining the fuel consumption. In the next chapter, the data sets used in this project will be introduced. On these data sets, we apply the Random Forest Regression model to determine the Feature Importance. Then the weights for the trip-level features are determined based on the feature rank, using which the weighted average is calculated and a final trip-score is provided. Figure 3.8 illustrates the final approach taken to design the scoring model.

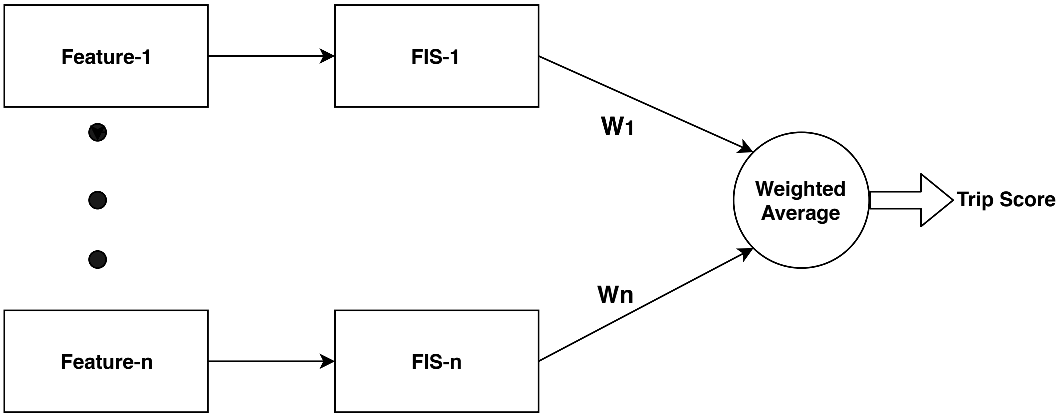


FIGURE 3.8: Trip-level scoring model

3.4 Design Decisions

A generic scoring algorithm takes into consideration the following attributes:

1. Target Value - A high score is provided if a variable/feature is able to meet the target value. For example, if a car is able to attain 70% of cruising in the whole trip it will be given a high score. Here, 70% is the target value.
2. Range and Sensitivity - Range typically sets certain limits to which a feature belongs. Sensitivity reflects the degree of change in input that affects the output.
3. Weight - To give relatively more importance to a feature than others, appropriate assignment of weights to a features are necessary.

We consider the above mentioned attributes and select fuzzy-logic to model the scoring algorithm. In fuzzy-logic the target value and range is set by defining the limits for the fuzzy sets. Slope of the fuzzy sets represent the sensitivity. But weights cannot be given using fuzzy logic. Therefore, we fuzzify and defuzzify every feature to obtain a single score for each parameter. By doing so, a quantitative feedback in the form of score is provided for every feature. In equation based scoring algorithms the target value and range have a definite limits. This requires expert knowledge and experience to set the limits, whereas by using fuzzy-logic these limits can be relaxed by overlapping the fuzzy sets. Also, the numerical ranges are represented linguistically, which makes it easier to reason and interpret. Multiple related features can be combined together to give a single score. Fuzzy logic is beneficial when it comes to approximation, vagueness and uncertainty. Overall, fuzzy-logic provides more flexibility in designing a scoring algorithm.

3.4.1 Design Decisions in Fuzzy Logic

The approach we have taken to design a scoring model is to fuzzify and defuzzify all the features. Another possible approach could be to fuzzify all the features at once and defuzzify to give a final score. The reasons for taking the former approach are:

- When two features were combined, the number of rules doubled in our case. If we were to combine, all the features the number of fuzzy rules would be very high.

- By fuzzifying all the features together, the priorities for each feature are equal.

Input fuzzy sets are overlapped to give a continuous range on defuzzification. Centre of Gravity is chosen as the defuzzification method because it provides a continuous variation by considering the area under the curve. Other defuzzification methods either provide a minimum or maximum value of the output fuzzy set.

3.5 Clustering

In the previous sections, feature extraction at the trip level based on the eco-driving rules were discussed. Along with feature extraction, a method to calculate score is also discussed. In this section, another technique is discussed to classify the drivers.

In this technique, no ground truth regarding the driving behavior is assumed and an unsupervised learning is performed on the features extracted from the data sets. Typically unsupervised learning consists of clustering where a cluster is visualized in 2 or 3 dimensional space. The features extracted cannot be visualized in this limited dimensional space, so there is a need to reduce the dimensions keeping implied data intact. This is done using dimensionality reduction techniques. The process of clustering is presented in Figure 3.9.

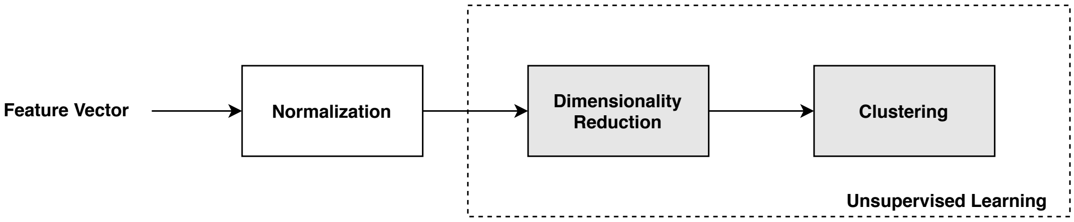


FIGURE 3.9: Clustering workflow.

Before reducing the number of dimensions, the features have to be standardized. This standardization process has to be carried out because the magnitude of extracted features varies largely. For example, average gear shift RPM is in the range of 1500-3500, whereas Cruise percentage varies from 0-100%. Although these features are different from each other, their varied magnitude can affect the classification. Dimensionality reduction techniques and clustering algorithms are sensitive to the magnitude of the features. We use standard scaler method for normalization as seen in Equation 3.7. Where μ and σ are mean and standard deviation of the feature x.

$$z = \frac{x - \mu}{\sigma} \tag{3.7}$$

Using PCA and t-SNE (discussed earlier in Chapter 2), nine features are reduced to two dimensions. Later, k-Means and Hierarchical clustering methods are used to categorize the driving behaviour. This leads to 4 different combinations, comprising of 2 dimensionality reduction techniques and 2 clustering algorithms. All these combinations will be tried to see which one brings out more meaning out of the data. To interpret the clusters formed from these combinations, the previously developed scoring algorithm is used to see if the scores among these clusters are unique.

3.6 Feature extraction for safe/unsafe driving

This section helps in establishing a relation between eco driving and safe driving. The features for safe / unsafe driving are extracted based on the research conducted in [16]. These features are based only on the speed as the source of the data was GPS signal. The features that are used to classify safe/unsafe trips are listed below:

- Maximum Speed
- Mean Speed
- Standard deviation speed
- Maximum acceleration
- Mean Acceleration
- Standard deviation acceleration
- Harsh Acceleration
- Harsh Braking

Previously, harsh acceleration and braking features were derived. The rest of the features mentioned above are statistical derivatives of speed and acceleration. These features will go through the clustering process described in the previous section. The clusters formed from these features will be compared with the clusters obtained for eco-driving behaviour. This comparison will help in concluding a relation between eco driving and safe driving.

3.7 Summary

This chapter focused on extracting various features which are important to analyze the driving behavior. Apart from the feature extraction, scoring model based on fuzzy logic was discussed. The design decisions made in the scoring algorithm were mentioned. The method of performing clustering on the features extracted and different clustering techniques were also mentioned in this chapter.

Chapter 4

Performance Evaluation

4.1 Data sets and Devices

In this section, we will explain the data sets used to model and evaluate the scoring and clustering process described earlier in Chapter 3. There are total three data sets available out of which one is an open source data set taken from [49]. The car used in this open source data set was a Seat Leon with a Diesel engine. The data was recorded at trip level and most of the routes consisted of trips between the German cities of Stuttgart, Reutlingen, Boblingen and Karlsruhe. The number of drivers who had participated in the acquisition of this data set are unknown. The trips were driven during the months of February, March, July and August. The device used to collect the data is a *PLX Kiwi 3* with the help of *OBD-II Doctor* smartphone application.

The second data set was collected using a Volkswagen Jetta with a Gasoline engine. All the parameters present in first data set along with an additional parameter viz. instantaneous fuel consumption are present in this data set. The car was driven by two drivers around the city of Orlando, Florida. Initially *PLX Kiwi 4* was used to collect the data as it had SD card logging capabilities. But due to technical failures, this device was replaced by a new OBD-II device viz. *OBD Link MX+*. Although, both the devices are capable of logging the data via a smartphone application using Bluetooth protocol, the preferred choice was *OBD Link MX+* as it was more reliable. Figure 4.1 and Figure 4.2 presents the devices used to log the data.

FIGURE 4.1: OBD Link MX+, The device used for data acquisition [50]

FIGURE 4.2: PLX Kiwi 4 [51]

The third data set was collected using a Peugeot 207 which was driven around the city of Enschede. This data set relatively consists of very less number of trips compared to the other data sets presented earlier.

Table 4.1 shows the detailed summary of the data sets. In the first data set DS_1 the parameter indicating fuel consumption was not present. Therefore, MAF was used to calculate the fuel consumption. The shorter trips (less than 10 km) were removed after calculating the fuel efficiency as they resulted in unrealistic values for fuel consumption. After filtering out the trips, there were a total of 64 trips left from this data set.

Data set	**DS_1**	**DS_2**	**DS_3**
Vehicle	Seat Leon	VW Jetta	Peugeot 207
Fuel Type	Diesel	Gasoline	Gasoline
Transmission	Manual	Automatic	Manual
# of Drivers	1	2	1
# of Trips	81	35	7
Total Distance (km)	4234	2297	110
Device	PLX Kiwi 3	OBD Link MX+	OBD Link MX+
Sampling Rate (Hz)	10	2	2

TABLE 4.1: Summary of the data sets

The second data set DS_2 was collected for evaluation and modeling purposes. In this data set, all the parameters along with the fuel consumption data were present. To differentiate between eco-driving and aggressive driving behavior, one of the driver was instructed to drive in an aggressive manner and other one to drive in rather non-aggressive manner (based on the theoretical driving rules presented earlier in Chapter 2). This instructed driving helped in setting the limits for the scoring algorithm.

The third data set DS_3 was collected to compare the scores and perform behavioural analysis among the drivers. The detailed descriptive statistics of the data sets are presented in Appendix A.

4.2 Evaluation Methods

The weights for the scoring algorithm are chosen based on the feature importance of different parameters. This feature importance is available in the ensemble machine learning algorithms. First, the evaluation metrics for ensemble learning method is presented. Then, the evaluation techniques for clustering algorithms are presented.

4.2.1 Performance Metrics for Ensemble Learning Models

We compare different ensemble learning algorithms with their default hyperparameters. The commonly used metric for regression algorithms is Mean Absolute Error (MAE). Further, MAE is converted into percentage as it is easier to conceptualize. Mean Absolute Percentage Error (MAPE) represents MAE in percentages. These metrics are robust to outliers, hence they are selected for measuring the performance of the models.

$$MAE = \frac{1}{n} \sum |actual - predicted| \tag{4.1}$$

$$MAPE = \frac{100\%}{n} \sum |\frac{actual - predicted}{actual}| \tag{4.2}$$

Based on the obtained accuracy's for different models, the model with highest accuracy is selected. Usually the default hyperparameters (settings of the machine learning model that can be tuned) don't give the highest accuracy, therefore, these values had to be tweaked. For this, a grid search with k-fold cross validation method was used. Using this method, the hyperparameters are searched for a large number of combinations.

4.2.2 Cluster Evaluation Techniques

A few clustering algorithms require the optimal number of clusters to be determined beforehand. One of them is k-Means clustering algorithm, where the number of clusters *k* has to be determined. The methods used for determining this optimal number are discussed below.

Silhouette Score

We measure the silhouette score for all the points in the cluster. Averaging the score over all of the clusters result in a certain value. A higher average silhouette score is preferred. This will be the first evaluation criteria to determine the number of optimal clusters.

Elbow Method

We determine the Within-Cluster Sum of Squares (WCSS) for *k* number of clusters. The optimal number of clusters is represented by the elbow point. If the elbow point cannot be decided from the plot, then we will use a combination of silhouette score to determine the optimal number of clusters.

Dendrogram

Hierarchical clustering methods form clusters in a hierarchical manner, which can be represented by a tree like structure called dendrogram. The roots of the dendrogram represent the trips and the vertical stems represent the distances between these trips. Using dendrogram, we can determine the number of clusters through visual inspection.

4.3 Results and Discussions

4.3.1 Feature Importance

Three ensemble learning models were evaluated to see which model provides the highest accuracy. The data set is splitted as 70% data for training and 30% data for testing. The input space comprises of nine features that were derived, and, the output space is fuel consumption. Out of the three (Random Forest, Gradient Boost, Extra Trees) regression models suggested, Random Forest performs slightly better than the other two models. Figure 4.3 shows the accuracy of the models on the two data sets. The results are shown for the default hyperparameters (baseline model).

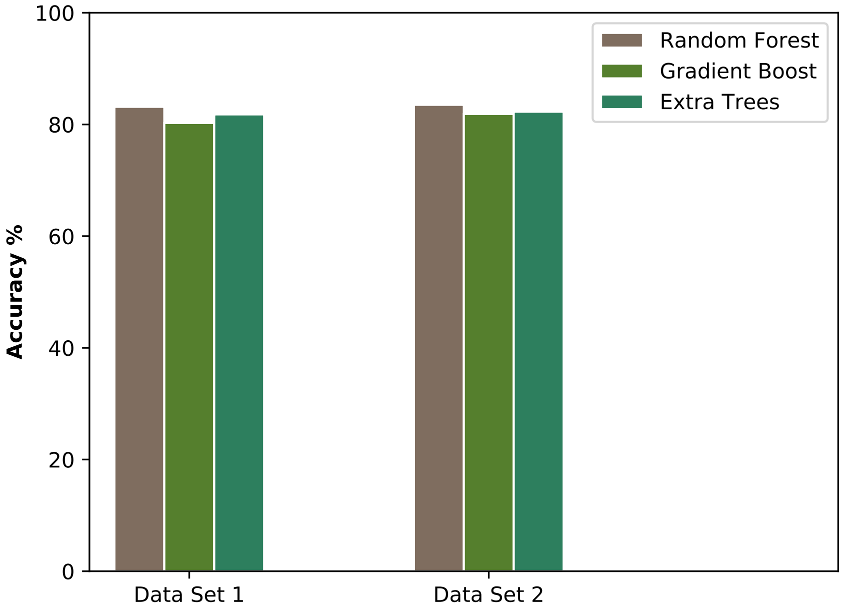


FIGURE 4.3: Performance of ensemble models for two sets

Random Forest model, due to better accuracy, was chosen to obtain the importance of the features. Random Forest model was tuned by performing 5-fold cross validation with Grid Search on both the data sets (DS_1 and DS_2). This method returns best parameters which further improves the accuracy of the model. Using these parameters the model accuracy improved by 0.9% for DS_1 and 0.83% for DS_2. Although an accuracy above 90% is preferred, the obtained results show that the data is good enough to evaluate the feature importance.

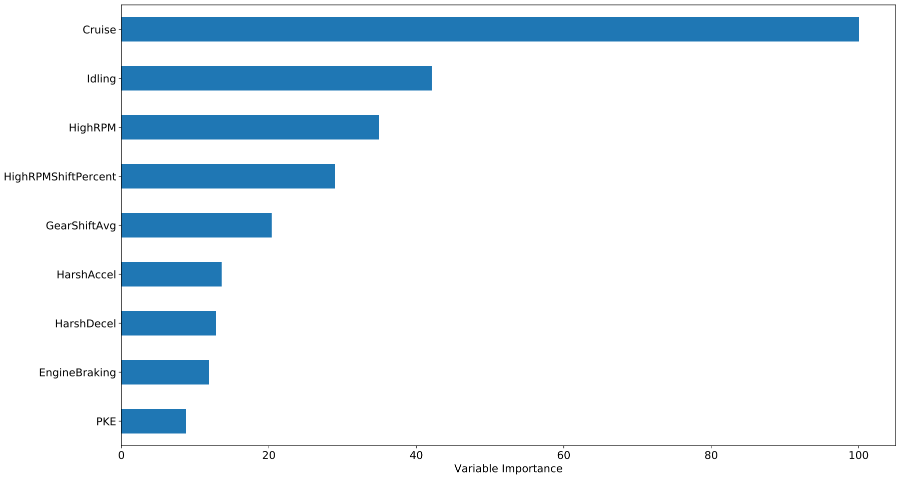


FIGURE 4.4: Feature Importance for data set DS_1

Now, the Random Forest model was used with the best parameters for each data set to get the feature importance. The plots for feature importance for DS_1 and DS_2 are as seen in Figures 4.4 and 4.5.

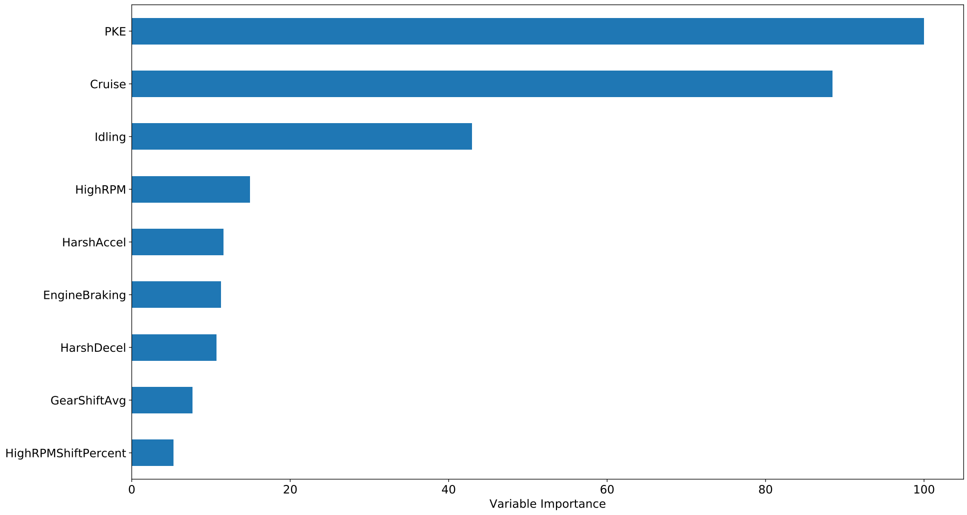


FIGURE 4.5: Feature Importance for DS_2

It can be seen from the feature importance of plot for DS_1 that *Cruise*, *Idling* and driving at *High RPM* are the most important features followed by *Gear Shifting*. One of the drawbacks of using feature importance property is that it shows reduced importance for correlated features [52]. To overcome this drawback, we evaluate feature importance of DS_2 and check the Pearson correlation between these features. The results are surprising as *PKE* which had the lowest importance shows up on the top of the list. Pearson correlation between Cruise and PKE is -0.67, which can be a possible reason for the drop in the rank of PKE. The rest of the features are nearly of similar importance.

The features related to gear shifting in DS_2 are ranked very low. This low ranking is due to the fact that the car which was used to generate data set DS_2 is automatic (no manual transmission system). The two aforementioned feature importance plots are used to decide the weights for each feature. The weights used in the scoring algorithm are as seen in Table 4.2.

Feature	Approximate Weight (%)
PKE	25
Cruise	20
Idling	15
High RPM	10
Gear Shift	10
Harsh Acceleration	7.5
Harsh Deceleration	7.5
Engine Braking	5

TABLE 4.2: Weights used in the scoring algorithm

The features PKE and Cruise account for 45% of the total weight. Idling is an independent feature and ranks higher in feature importance plots, therefore it is given a higher value than the rest. To decide the weight for the rest of the features the correlation analysis is applied with reference to the feature importance. Gear Shift and High RPM are equally correlated with fuel efficiency, hence an equal weight is assigned. Harsh Acceleration and Deceleration are equally correlated with each other and show a high correlation with PKE. Therefore, the weights given are comparatively low and equal. Engine braking shows a positive correlation with fuel efficiency, which was unexpected. The reason for this is that the trips which were driven aggressively have higher fluctuation in speed. So, this results in an increased braking naturally. To minimize the adverse effects of engine braking on the score, a low value is assigned. Having decided the weights in the scoring algorithm, next step is to validate the scoring algorithm.

4.3.2 Validation of the Scoring Algorithm

Validation of the scoring algorithm is done against the fuel consumption as it is the main dependent variable for eco-driving. Data worth 16 trips was collected separately with Volkswagen Jetta. The car was driven by two drivers and different routes (Highway and City) were taken. Some of the trips are driven while following eco-driving rules, whereas the others are driven aggressively. The fuel consumption was directly taken from the vehicle's fuel gauge (display) and after every trip the fuel gauge was reset.

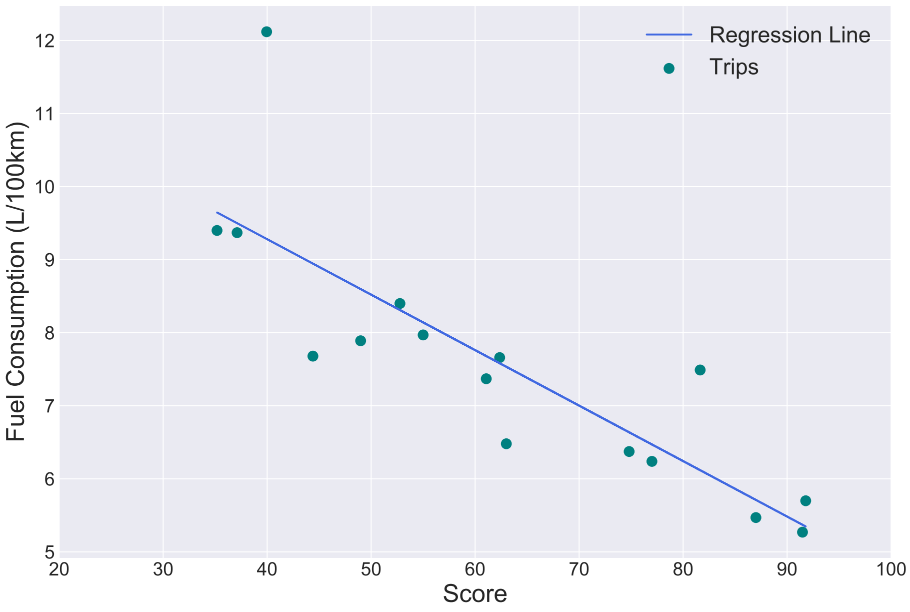


FIGURE 4.6: Score vs Fuel Consumption

From the Figure 4.6 it can be observed that the score and fuel consumption are highly correlated. To quantify the relationship between the two variables, we perform Pearson Correlation analysis. It is found that the two variables have a correlation magnitude of -0.84, which is quite high in the range of -1 to 1. A negative correlation indicates that a higher score results in lesser fuel consumption. This way the scoring algorithm is validated against the fuel consumption.

4.3.3 Comparing Drivers based on the Scoring Algorithm

The three data sets described earlier in Table 4.1 comprises of the driving data of four drivers. Data set DS_2 comprises of two drivers, the rest of the two data sets are driven by one driver. We compared the average and standard deviation of the scores to observe if consistent driving is the key characteristic of certain drivers.

Analysis is performed on the drivers and results are tabulated in Table 4.3. From this table, it is observed that driver B and D are the most consistent drivers because the standard deviation of their score is low. The average score for driver D is slightly low due to the city driving conditions. In city driving conditions, the cruising behavior is reduced and idling behavior is increased which contributes to a lower score. Driver A's trips mostly are driven on the German highways, and these trips are not very efficient because certain trips are driven at very high engine speeds. Driver C was instructed to driver aggressively to check the model behavior. Generation of lowest score among all other drivers indicate that aggressive behavior is captured by the model.

Data set	**Driver**	**Average Score**	**Std. Deviation of Score**	**Remark**
DS1	A	66	8	Inefficient - Highway driving
DS2	B	85.24	4.67	Efficient driver
	C	49.15	9.15	Aggressive driver
DS3	D	74.3	3.8	Consistent driver - City driving

TABLE 4.3: Aggregated trip-level scores of the drivers

From the data sets, it is observed that the drivers have driven their vehicle at varying speeds. This might be because of the different speed limits imposed by the authorities at different places. The highest speed limit in Florida is 80 miles per hour (128 km/hr). The maximum speed of the drivers in DS_2 is 126 km/hr, which is well within the speed limits of Florida state. Since, German highways do not have speed limits, the maximum speed achieved by driver A is 217 km/hr. The maximum speed achieved by driver D is 96 km/hr, this might be due to the suburban speed limits in the Netherlands. Overall, the drivers B and D are efficient and consistent compared to the other drivers.

To test and verify the generalizability of the scoring algorithm on trips that are not present in the data set, an experiment was conducted. In this experiment, two drivers participated and the car used was a Renault Megane. The chosen route was from Enschede-Cologne for the first driver and Cologne-Enschede for the second driver. This experiment also served as a means to portray the gamification concept of the scoring algorithm developed.

From the Figure 4.7 it is observed that first driver is more efficient than the second driver. Fuel consumption obtained from the car's dashboard also verifies this fact. The comparison of score ad fuel consumption is presented in the Table 4.4. The car's dashboard displayed the distance traveled without fuel consumption, which is basically engine braking. The total amount of engine braking for the entire trip was 8%, which was calculated as a part of feature extraction. The car's display showed that the distance traveled without fuel consumption as 15 km for a total trip length of 192 km, which amounts to 7.8%. This way the feature extraction for engine braking is verified.

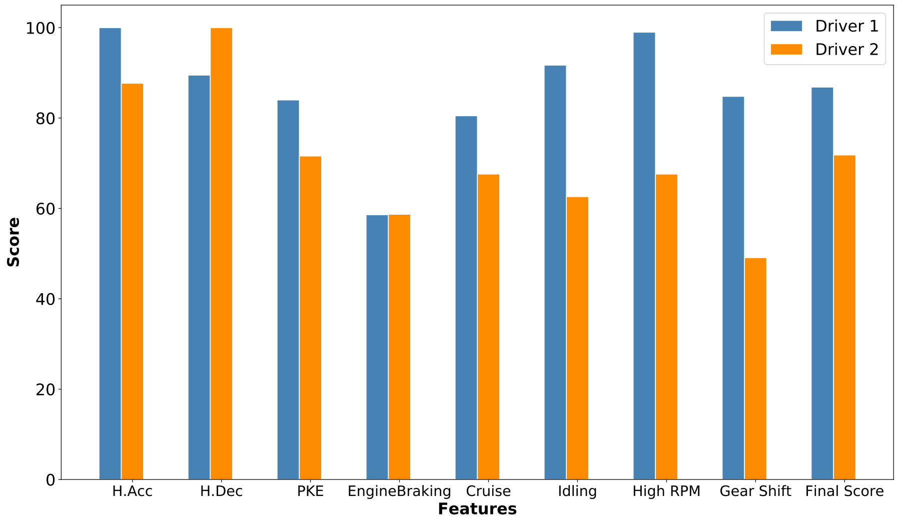


FIGURE 4.7: Score comparison of the two drivers

Driver	Score	Fuel Consumption (L/100km)
1	87	5.1
2	72	6.5

TABLE 4.4: Score and fuel consumption for Enschede-Cologne trip

4.3.4 Clustering

We use the second data set (DS_2) to perform clustering as it is more diverse in terms of driving behaviour. First, we reduce the number of dimensions using PCA and t-SNE. Then, we perform clustering using k-Means and agglomerative methods. Later, we compare the different methods and determine which method is more applicable for this data set. The clusters obtained are evaluated with the scores obtained from the scoring algorithm.

4.3.4.1 Principal Component Analysis

By applying PCA on the data set, the number of features are reduced to a few principal components. The number of components is decided based on the scree plot. In this analysis, highly correlated features are transformed into linearly uncorrelated features. From Figure 4.8, it is clear that the two principal components PC1 and PC2 obtained account for more than 80% of the variance. It means that the nine features can be represented with these two principal components.

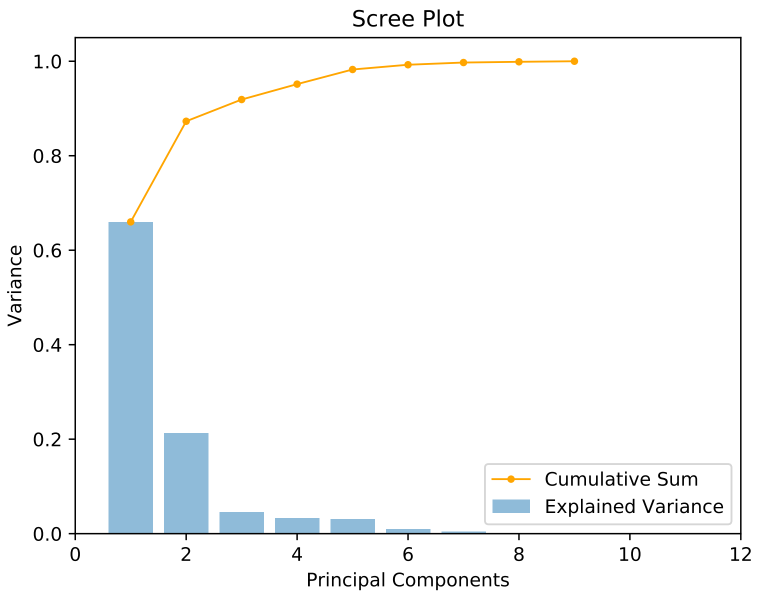


FIGURE 4.8: Scree plot showing the variance of principal components

The summary of principal components of the data set are shown in Table 4.5. In PC1, cruise and PKE have the largest coefficients, whereas, in PC2, idling and gear shift features have the largest coefficients. The coefficients with large magnitudes will determine the position of the trips on a two dimensional projection.

Feature	PC1	PC2
Harsh Acceleration	0.32	0.31
Harsh Deceleration	0.29	0.45
PKE	0.39	0.07
Engine Braking	0.36	-0.21
Cruise	-0.39	-0.13
Idling	0.28	0.41
High RPM	0.30	-0.35
Gear Shift Average	0.31	-0.41
% of High RPM Shifts	0.29	-0.40

TABLE 4.5: Summary of the component loadings

4.3.4.2 T-Stochastic Neighbour Embedding

t-SNE is a non-linear dimensionality reduction technique mostly used for visualization purpose. One of the advantage of using this technique is that it brings similar points in higher dimensional space closer when projected onto the lower dimensional space. To perform t-SNE one has to consider tuning a few important parameters. They are : number of components required, perplexity and number of iterations to converge. For this data set, the number of components to keep is set to 2, as we want to visualize in two dimensions. Perplexity is set to 14, and the number of iterations is set to 10000. These parameters are set by trial and error experimentation [53].

4.3.4.3 k-Means Clustering

First we perform k-means clustering with PCA reduced population, then we will experiment with t-SNE. For k-means clustering, the k value has to be determined. We use silhouette score and elbow method to arrive at this number.

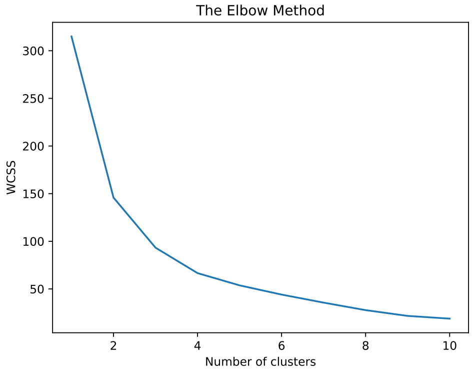


FIGURE 4.9: Within-cluster sum of squares for PCA reduced population with k-means clustering

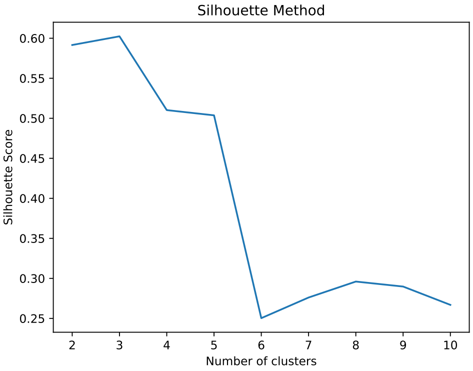


FIGURE 4.10: Silhouette scores for PCA reduced population with k-means clustering

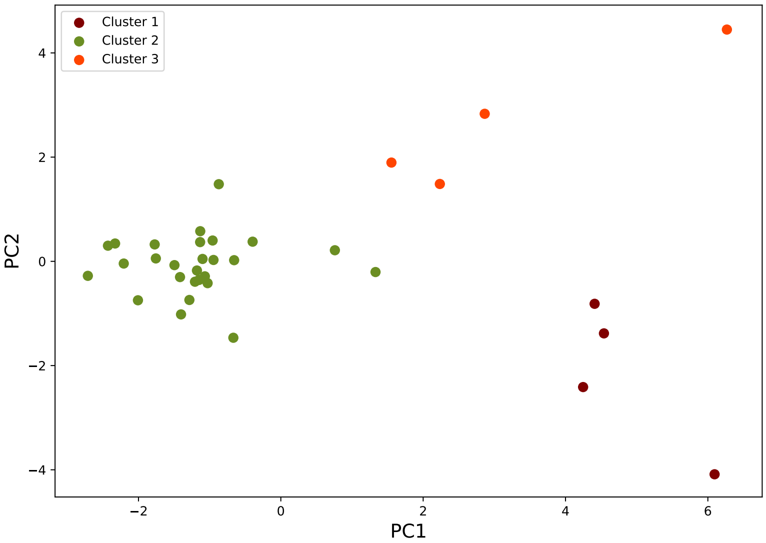


FIGURE 4.11: PCA projection of k-means clustering

From the Figure 4.9 the elbow point is not very clear as it can be 2, 3 or 4. Therefore, elbow method is not considered. In the Figure 4.10 the highest score is obtained when the number of clusters are 3. So, the optimal number of clusters from these plots are considered to be 3.

Figure 4.11 shows the clusters obtained by reducing the dimensions through PCA. Among the three clusters obtained, cluster 2 represents the trips driven efficiently. The other two clusters are driven not so efficiently and cluster 3 has the most aggressive trips. The trips in cluster 3 are driven at high RPM and have a higher harsh deceleration count compared to cluster 2.

Next, we reduce the features using t-SNE and cluster with k-Means. Figure 4.12 and 4.13 show the cluster evaluation methods. The elbow point drops when the number of clusters are 2 and 4, which makes it ambiguous. Silhouette score is highest when the cluster number is 2. Therefore, the number of optimal clusters is considered to be 2.

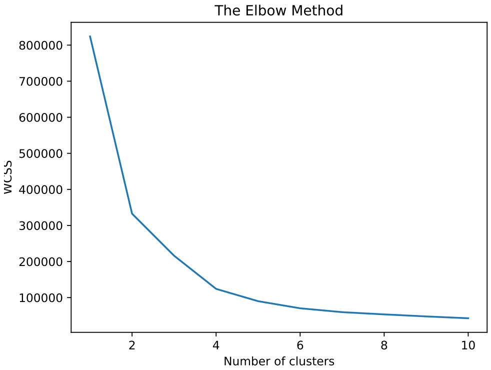


FIGURE 4.12: Within-cluster sum of squares for t-SNE reduced population with k-means clustering

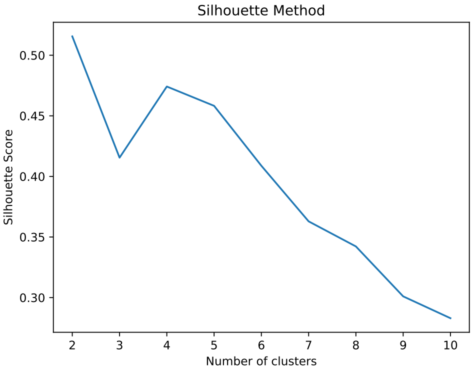


FIGURE 4.13: Silhouette scores for t-SNE reduced population with k-means clustering

The clusters obtained from t-SNE projection of k-means is as seen in Figure 4.14. The obtained clusters clearly differentiate the driving behaviour. Cluster 1 represents the trips driven in an inefficient way. Cluster 2 represents the trips that are driven in a very efficient manner. From this plot a clear distinction can be made between the driving behaviour.

Cluster	PCA	t-SNE
1	49.6	49.1
2	83.5	85.2
3	42.3	-

TABLE 4.6: Average scores of the clusters obtained by k-means clustering

To interpret the clusters, we use the trip-level score as the ground truth. Table 4.6 shows the average scores of the trips obtained by k-means clustering. The scores between the clusters are significantly different. In the previous section, we have shown the correlation between fuel consumption and trip-score. Therefore, if the score for a given cluster is high, it indicates that the fuel efficiency is high which is a characteristic of eco-driving.

4.3.4.4 Hierarchical Clustering

Hierarchical clustering can be performed in two ways : agglomerative and divisive. When the number of data points are less, the choice to be made is agglomerative

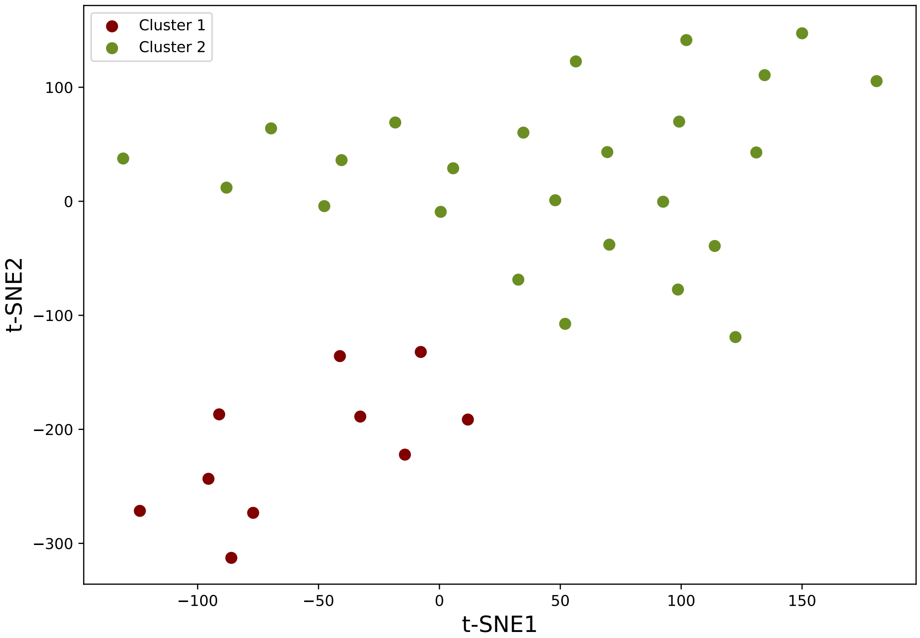


FIGURE 4.14: t-SNE projection of k-means clustering

clustering. Further, the linkage criteria and the dissimilarity measure has to be specified for performing clustering. We choose *ward* linkage criteria and *euclidean* distance measure for performing clustering, as these are the commonly used parameters. To decide the the number of clusters, we visualize the dendrogram of the PCA and t-SNE projections. The length of the vertical lines of the dendrogram represent the dissimilarity between clusters. The longer the lines, more dissimilar are the clusters.

From figure 4.15 and 4.16 it is observed that the the lines colored in blue have the largest vertical distance. This is used to determine the number of clusters. From both PCA and t-SNE projections it clear that two clusters can be formed. We specify this parameter to perform agglomerative clustering and then project the results on a scatter plot.

From the Figures 4.17 and 4.18 two clusters are obtained. Among the two clusters, one of them is driven efficiently and the other one is driven aggressively. The clusters formed using PCA and t-SNE have the exact same trips, which validates the clusters obtained and the dimensionality reduction techniques. Table 4.7 shows the average scores of the clusters, it is observed that the scores among the clusters are equal.

Cluster	PCA	t-SNE
1	49.1	85.2
2	85.2	49.1

TABLE 4.7: Average scores of the clusters obtained from agglomerative clustering

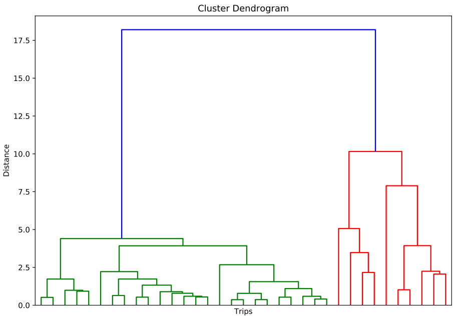


FIGURE 4.15: Dendrogram of PCA reduced population with agglomerative clustering

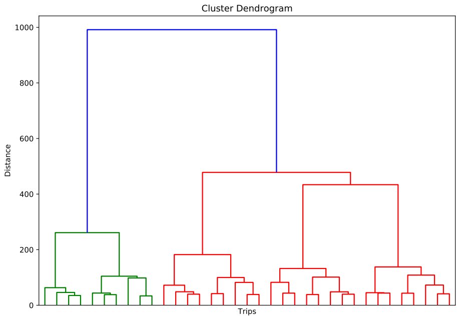


FIGURE 4.16: Dendrogram of t-SNE reduced population with agglomerative clustering

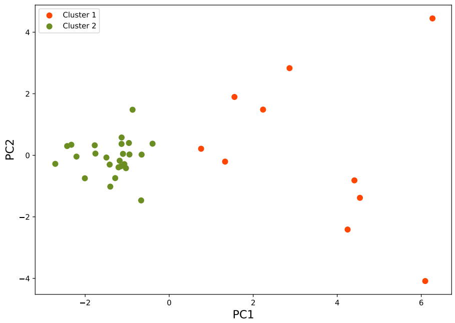


FIGURE 4.17: PCA projection of agglomerative clustering

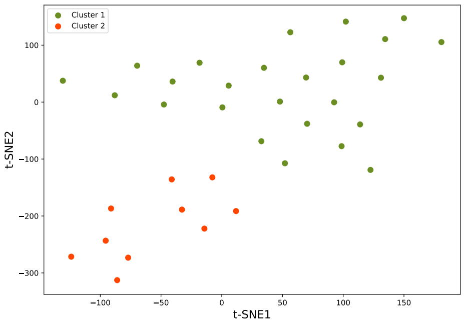


FIGURE 4.18: t-SNE projection of agglomerative clustering

4.3.4.5 Comparing PCA and t-SNE

Although we had very few trips in our data set, PCA and t-SNE both were applicable to perform dimensionality reduction. PCA provides clear information about the number of components, their component loadings and separates correlated variables. Regarding t-SNE, it doesn't provide any information about it's lower dimensional components. Hence, it has to be treated with extra attention because of it's black-box nature. Having the ground truth in the form of scores of the trips helps making t-SNE interpretation simpler.

There were no inputs that are required to be provided to the PCA algorithm for performing dimensionality reduction, but to get a good cluster representation using t-SNE, several hyperparameters have to be tuned. The hyperparameter tuning for t-SNE is more of a guess work. Comparing the clusters obtained visually, t-SNE clusters are compactly grouped compared to the PCA. This means, the local structure of the data points (trips) are well preserved using t-SNE.

4.3.4.6 Comparing K-means and Agglomerative Clustering

K-means clustering requires the number of initial clusters to be specified. For this, we had to perform cluster evaluation techniques before performing the actual clustering. Agglomerative clustering groups the trips without the need of any input. Using the dendrogram, the optimal number of clusters can be decided. To compare the best performing combination of the dimensionality reduction technique and clustering algorithm we use silhouette score for assessment. From Table 4.8, it observed that the clusters obtained from the combination of k-means and PCA has the highest silhouette score. Clustering on PCA reduced population has a better score than the t-SNE reduced population.

	k-means	**Agglomerative**
PCA	0.60	0.55
t-SNE	0.51	0.51

TABLE 4.8: Silhouette scores for the combination of dimensionality reduction techniques and clustering algorithms

Overall, the combination of clustering and dimensionality reduction techniques provide a good classification of driving behaviour from the trip-level features. This also means that the feature vectors that were derived are a good representation of the eco-driving behaviour. This is also validated against the scores generated at the trip-level from the scoring algorithm.

4.3.5 Evaluation of external factors on eco-driving

Route choice and weather affects the driving behaviour. To show the effects of route choice, a trip was planned from point A to B using two routes in the city of Orlando, Florida. The first route passes through the city, and the second route takes a highway. The route taken through the city had 14 stops (due to traffic) and the highway route had 4 stops. The distances for the city and highway were 46 and 50 kilometres respectively. Both the trips were driven by a single driver in an efficient manner.

The route taken through the city resulted in a fuel efficiency of 6.1 L/100km, whereas, the other route resulted in a fuel efficiency of 5.2 L/100km. Due to the traffic (stop-lights), idling was higher by 7%, cruising reduced by 15% and positive kinetic energy increased by 40% in comparison to the trip which took the highway route. The overall improvement of 14% in fuel efficiency was observed. Since, the scoring algorithm is designed based on the parameters that affect the fuel consumption, the trip driven through the city received a score of 73 and the trip driven by the highway received a score of 88. Through this experiment, it was inferred that apart from the driving behaviour route choice also influences the eco-driving.

The trips in the data set DS_1 presented earlier are driven in Germany. These trips have the date and location of the journey recorded. Using these labels, the historical weather information [54] was gathered to analyze the effects of weather in the driving behavior. The weather data includes temperature, humidity, wind speed and precipitation. The city at which the trip started is taken as the reference for the weather data. To see the relation between weather and the driving behaviour, correlation analysis was performed. After this, trip-level scores were checked to see any effects of weather on the driving behavior.

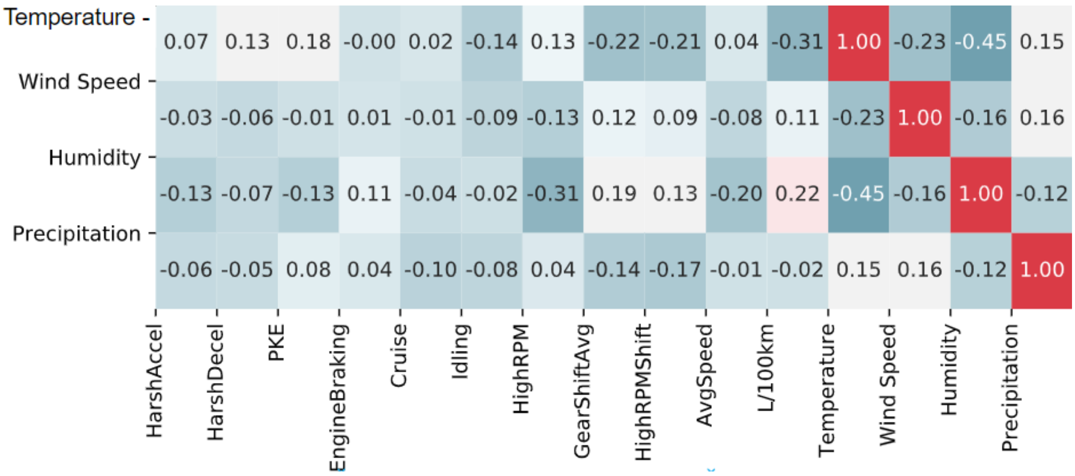


FIGURE 4.19: Correlation of eco-driving features and weather

From the correlation map in the Figure 4.19 it can be observed that most of the weather parameters have very less correlation with the driving features. However, the humidity and High RPM features show a correlation of -0.31. It can be inferred that when the humidity is low (good weather), the driver seems to drive aggressively at higher engine speeds. The average speed also shows a positive correlation of 0.2 with the humidity indicating that the driver chooses to drive at higher speeds. Research conducted in [55] shows that the drivers display aggressive behaviour with the increase in temperature. The temperature and humidity show a correlation of -0.45 which means that the decrease in humidity results in the increase of temperature. So, temperature and humidity are linked to the driver's higher engine speeds.

The scores in a given calendar month didn't show significant difference on occasions when there was heavy precipitation. The amount of data that is used to perform driving analysis with respect to weather is considerably less. Historical weather information that is extracted is also subject to the weather station that provides the data. Having accurate location and larger data set might have provided better results.

4.3.6 Safe driving and Eco-driving

The parameters that were described in the previous chapter for characterizing safe / unsafe driving are used to perform clustering. From the eco-driving features, harsh acceleration and deceleration are removed as these are present in the safe/unsafe feature set. The average silhouette scores for a combination of dimensionality reduction techniques and clustering algorithms are calculated. The results obtained are displayed in Table 4.9. PCA produced a higher score for safe / unsafe feature set due to the less number of trips which resulted into scattered data points. The performance of t-SNE is nearly the same in both the feature sets. Also, agglomerative clustering performs slightly better than k-means method. Therefore, t-SNE and agglomerative clustering were used to obtain the clusters. From the resulting clusters, analysis was performed on the overlap of safe and unsafe driving against eco-driving trips.

From the silhouette scores the optimal number of clusters that can be formed are 2 for both the feature sets. The results obtained from clustering on the two feature sets are as seen in Figure 4.20 and 4.21. It is observed that the cluster 1 (green) in both the figures score higher than cluster 2 (red). The scores for the clusters are presented in Table 4.10. The average score for cluster 2 even after removing two features are the same compared to the one obtained in Table 4.7.

Technique		Safe/Unsafe features	Eco-driving features
PCA	k-means	0.60	0.51
	Agglomerative	0.60	0.51
t-SNE	k-means	0.53	0.54
	Agglomerative	0.54	0.54

TABLE 4.9: Silhouette scores for a combination of dimensionality reduction and clustering algorithms

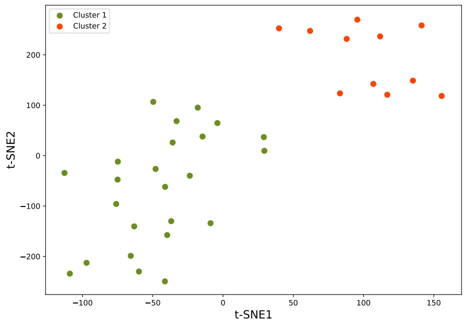


FIGURE 4.20: Clustering on safe/unsafe driving features

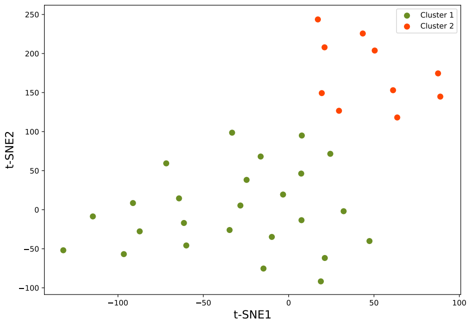


FIGURE 4.21: Clustering on eco-driving features

Cluster	**Safe/Unsafe**	**Eco-driving**
1	85.58	85.25
2	51.71	49.15

TABLE 4.10: The scores of the clusters obtained from two feature sets

By comparing the cluster labels of each trip we observe that only one trip from cluster 1 of safe/unsafe feature set does not belong to the cluster 1 of eco-driving feature set. This resulted into a very high overlap of 97% between safe driving and eco-driving.

The average values of the two feature sets (safe/unsafe and eco-driving) are seen in Table 4.11. It is observed that the standard deviation of the acceleration is higher in cluster 2 compared to the cluster 1. Also, the features harsh acceleration and deceleration have very high average values in cluster 2 for safe/unsafe feature set. In the feature set related to eco-driving, PKE is significantly high in the cluster 2. Also, the cruising, high RPM and features related to gear shift are higher. So, cluster 2 among both the feature sets are considered as unsafe and inefficient trips.

Safe/Unsafe			Eco-driving		
Feature	**Cluster 1**	**Cluster 2**	**Feature**	**Cluster 1**	**Cluster 2**
Harsh Acc.	2.92	23.6	PKE	1455	4995
Harsh Dec.	4.11	18.9	Cruise	72.4	31
Mean Speed	86.4	61.9	Idling	6.1	14
Max Speed	122	122	High RPM	0	3.86
Std. Speed	33	35	Avg. Gear Shift	2069	2487
Max. Acc.	14.16	21.63	% High RPM Shift	14	36.5
Std. Acc.	1.76	2.69	Engine Braking	5.8	19.05

TABLE 4.11: Averages of two feature sets among the clusters

4.4 Summary

This chapter presented the results of the methods discussed earlier in Chapter 3. Firstly, the results obtained from the Random Forest's feature importance were presented. Secondly, the scoring algorithm is validated against fuel consumption. Then, the different drivers among the data sets were compared. Third, the unsupervised learning was performed and different methods were compared. Fourth, the effects of external factors on eco-driving behaviour were presented. Lastly, the correlation between safe-driving and eco-driving results are presented.

Chapter 5

Conclusion

This thesis work presented a detailed description of the parameters which are obtained from the vehicle's ECU. Feature extraction process along with the scoring model based on fuzzy logic were also described in the earlier chapters. This scoring model helps in classification of driver behavior as economic (eco) or aggressive. Another technique which is based on clustering was also introduced to perform the classification of driving behavior.

During the literature survey, important eco-driving rules were noted down. Based on these rules, nine features were extracted from the time-series data collected from the vehicle. On these extracted features, random forest model is used which evaluates the importance of features for characterizing eco-driving behavior. From this analysis, it was observed that positive kinetic energy (PKE), cruising, idling and high RPM driving are the important features which best characterize the eco-driving behavior. Also, there are some parameters like engine braking which had very less correlation with the fuel consumption, and thus do not contribute much to the eco-driving behavior.

To perform the quantitative evaluation of driving behavior, a fuzzy logic based scoring model was developed. This model makes use of all the extracted features from the data set. The scoring model is validated by checking the correlation with fuel consumption. Based on this analysis, it was observed that calculated score and the fuel consumption were highly correlated. This analysis helps in concluding that the scoring model works well to classify the eco-driving behavior. Using this scoring algorithm, the drivers among the data sets were compared. It was found that some drivers are consistent and efficient, while the others are aggressive. This aggressive behavior is also reflected in the scoring model.

Apart from the scoring model, another technique based on unsupervised machine learning concepts was used to classify the driving behavior. Design space was explored for two dimensionality reduction techniques and two clustering algorithms. Based on this exploration, a combination of PCA (dimensionality reduction technique) and K-means (clustering technique) provided better results. These results were based on their silhouette score. From the results obtained, it can be concluded that the combination of dimensionality reduction techniques and clustering algorithms can distinguish the driving behaviour clearly. Among the clusters that were obtained, at least one of them represented eco-driving behavior which indicates that machine learning algorithms can be employed for driving behavior classification.

Literature survey was performed to identify the features that relate to safe and unsafe driving. An analysis was performed using the clustering algorithm based on these features. The results obtained from this analysis were compared to the analysis made for eco-driving. These results showed an overlap of 97% between eco-driving and safe-driving. This observation concludes that safe-driving and eco-driving are highly correlated to each other.

Data sets were also analyzed to check the effect of route choice and weather on the driving behavior and fuel consumption. Based on the limited data set, very less correlation was found between weather and driving behavior. It was also observed that route choice affects the fuel consumption and thus the eco-driving behavior.

5.1 Future Work and Recommendation

The results and conclusions presented in this thesis are based on the limited time frame available for this work. Due to this, there are a few tasks which were left unattended, and, few tasks which required more inputs in terms of data sets and time. Following points present future work and recommendations if this thesis work is carried out further :

- Collection of more data to tune the limits and weights present in the scoring algorithm.
- Identification of more features to make the scoring model more accurate.
- Integration of scoring model into a smartphone application or vehicle to give real-time feedback to the drivers about their driving behavior.
- Currently, the scoring model does not differentiate between city and highway driving. Provisions should be made in the model to identify these different scenarios and adjust the model accordingly for accurate results.
- Cruising speed is vehicle dependent. Implementation of this scoring model based on vehicles specifications can lead to better cruising scores.
- Contextual information like road conditions and traffic information can be included to accurately analyze the driving behavior.
- Currently, the scoring model is meant for cars. This scoring model can further be extended to analyze the driving behavior on buses and long-haul trucks.
- Real-time feedback generation should be added to the model to make the system more informative.
- Cars tested in this work had a standard transmission. This model should be extended to account for different types of transmissions like Continuous-Variable-Transmission (CVT).

Bibliography

[1] *Annual Global Road Crash Statistics*. URL: https://www.asirt.org/safe-travel/road-safety-facts/ (visited on 05/20/2019).

[2] Tatsuaki Osafune et al. "Analysis of accident risks from driving behaviors". In: *International journal of intelligent transportation systems research* 15.3 (2017), pp. 192–202.

[3] *Vehicular Pollution*. URL: http://www.pollutionissues.com/Ve-Z/Vehicular-Pollution.html (visited on 05/20/2019).

[4] *Health and Sustainable Development*. URL: https://www.who.int/sustainable-development/transport/health-risks/climate-impacts/en/ (visited on 05/20/2019).

[5] *Benefits of Eco-driving*. URL: http://www.together-eu.org/docs/102/TOGETHER_Eco-driving_5_Handout_15.pdf (visited on 05/20/2019).

[6] *Chevin Fleet Management*. URL: https://www.chevinfleet.com/au/news/what-is-fleet-management-software/ (visited on 10/06/2019).

[7] Yuhan Huang. "Eco-driving Technology for Logistics Transport Fleet to Reduce Fuel Consumption and Emissions". In: *UTS* (2018).

[8] Michael Sailer et al. "How gamification motivates: An experimental study of the effects of specific game design elements on psychological need satisfaction". In: *Computers in Human Behavior* 69 (2017), pp. 371–380.

[9] Dominik L Schall, Menas Wolf, and Alwine Mohnen. "Do effects of theoretical training and rewards for energy-efficient behavior persist over time and interact? A natural field experiment on eco-driving in a company fleet". In: *Energy Policy* 97 (2016), pp. 291–300.

[10] Maria Zarkadoula, Grigoris Zoidis, and Efthymia Tritopoulou. "Training urban bus drivers to promote smart driving: A note on a Greek eco-driving pilot program". In: *Transportation Research Part D: Transport and Environment* 12.6 (2007), pp. 449–451.

[11] Cindie Andrieu and Guillaume Saint Pierre. "Comparing effects of eco-driving training and simple advices on driving behavior". In: *Procedia-Social and Behavioral Sciences* 54 (2012), pp. 211–220.

[12] M Jensen, J Wagner, and K Alexander. "Analysis of in-vehicle driver behaviour data for improved safety". In: *International journal of vehicle safety* 5.3 (2011), pp. 197–212.

[13] Peter Händel et al. "Insurance Telematics: Opportunities and Challenges with the Smartphone Solution." In: *IEEE Intell. Transport. Syst. Mag.* 6.4 (2014), pp. 57–70.

[14] Zhishuo Liu, Qianhui Shen, and Jingmiao Ma. "A driving behavior model evaluation for UBI". In: *International Journal of Crowd Science* 1.3 (2017), pp. 223–236.

[15] M Jasinski and F Baldo. “A Method to Identify Aggressive Driver Behaviour Based on Enriched GPS Data Analysis”. In: *The Ninth International Conference on Advanced Geographic Information Systems, Applications, and Services* (2017).

[16] Jeroen Bijman. “Cluster driving behaviour and assigning clusters to safe and unsafe driving behaviour through raw GPS trajectory data”. MA thesis. Tilburg University School of Humanities, 2017.

[17] *The Golden Rules of Eco-driving*. URL: http://www.ecodrive.org/en/what_is_ecodriving/the_golden_rules_of_ecodriving/ (visited on 11/08/2019).

[18] Gary A Davis, Sujay Davuluri, and Jian Ping Pei. “A Case Control Study of Speed and Crash Risk, Technical Report 3: Speed as a Risk Factor in Run-off Road Crashes”. In: (2006).

[19] Charles M Farmer. “Relationship of traffic fatality rates to maximum state speed limits”. In: *Traffic injury prevention* 18.4 (2017), pp. 375–380.

[20] Xiaoyu Zhu, Xianbiao Hu, and Yi-Chang Chiu. “Design of Driving Behavior Pattern Measurements Using Smartphone Global Positioning System Data”. In: *International Journal of Transportation Science and Technology* 2.4 (2013), pp. 269–288.

[21] *Speed Kills MPG*. URL: https://www.mpgforspeed.com/ (visited on 05/14/2019).

[22] Jongwoo Choi et al. “Analysis of the vehicle information for the economical and the ecological driving pattern”. In: *2009 11th International Conference on Advanced Communication Technology*. Vol. 1. IEEE. 2009, pp. 867–869.

[23] Chi-pan Hwang et al. “Apply Scikit-Learn in Python to Analyze Driver Behavior Based on OBD Data”. In: *2018 32nd International Conference on Advanced Information Networking and Applications Workshops (WAINA)*. IEEE. 2018, pp. 636–639.

[24] Fatjon Seraj et al. “A smartphone based method to enhance road pavement anomaly detection by analyzing the driver behavior”. In: *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers*. ACM. 2015, pp. 1169–1177.

[25] Jin-Hyuk Hong, Ben Margines, and Anind K Dey. “A smartphone-based sensing platform to model aggressive driving behaviors”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM. 2014, pp. 4047–4056.

[26] Tanushree Banerjee, Arijit Chowdhury, and Tapas Chakravarty. “MyDrive: Drive behavior analytics method and platform”. In: *Proceedings of the 3rd International on Workshop on Physical Analytics*. ACM. 2016, pp. 7–12.

[27] Chen Chen et al. “Driver’s Eco-Driving Behavior Evaluation Modeling Based on Driving Events”. In: *Journal of Advanced Transportation* 2018 (2018).

[28] Javier E Meseguer et al. “Drivingstyles: a mobile platform for driving styles and fuel consumption characterization”. In: *Journal of Communications and networks* 19.2 (2017), pp. 162–168.

[29] Branislav Sarkan et al. “Vehicle fuel consumption prediction based on the data record obtained from an engine control unit”. In: *MATEC Web of Conferences*. Vol. 252. EDP Sciences. 2019, p. 06009.

[30] *Is Petrol Usage Proportional to Gas Pedal Position*. URL: https://mechanics.stackexchange.com/questions/11832/is-petrol-usage-proportional-to-gas-pedal-position (visited on 05/16/2019).

[31] João C Ferreira, José de Almeida, and Alberto Rodrigues da Silva. "The impact of driving styles on fuel consumption: A data-warehouse-and-data-mining-based discovery process". In: *IEEE Transactions on Intelligent Transportation Systems* 16.5 (2015), pp. 2653–2662.

[32] VAM Heijne, NE Ligterink, and U Stelwagen. "Effects of Driving Behaviour on Fuel Consumption". In: *Poster session presented at the Transport Air Pollution Conference, Zürich, Switzerland*. 2017.

[33] Víctor Corcoba Magaña and Mario Muñoz-Organero. "Artemisa: A personal driving assistant for fuel saving". In: *IEEE Transactions on Mobile Computing* 15.10 (2015), pp. 2437–2451.

[34] N Zacharof et al. "Review of in use factors affecting the fuel consumption and CO2 emissions of passenger cars". In: *European Commission* (2016).

[35] Karsten Jakobsen, Sabrine CH Mouritsen, and Kristian Torp. "Evaluating eco-driving advice using GPS/CANBus data". In: *Proceedings of the 21st ACM SIGSPATIAL international conference on advances in geographic information systems*. ACM. 2013, pp. 44–53.

[36] Zoran Constantinescu, Cristian Marinoiu, and Monica Vladoiu. "Driving style analysis using data mining techniques". In: *International Journal of Computers Communications & Control* 5.5 (2010), pp. 654–663.

[37] R. Massoud et al. "Eco-driving Profiling and Behavioral Shifts Using IoT Vehicular Sensors Combined with Serious Games". In: *2019 IEEE Conference on Games (CoG)*. 2019, pp. 1–8. DOI: 10.1109/CIG.2019.8847992.

[38] Luigi Fortuna et al. *Soft computing: new trends and applications*. Springer Science & Business Media, 2001.

[39] Ye Ren, Le Zhang, and Ponnuthurai N Suganthan. "Ensemble classification and regression-recent developments, applications and future directions". In: *IEEE Computational intelligence magazine* 11.1 (2016), pp. 41–53.

[40] Leo Breiman. "Random forests". In: *Machine learning* 45.1 (2001), pp. 5–32.

[41] Xin Jin and Jiawei Han. *K-Means Clustering In: Sammut C, Webb GI, editors. Encyclopedia of Machine Learning*. 2010.

[42] Fionn Murtagh and Pedro Contreras. "Methods of Hierarchical Clustering". In: *CoRR* abs/1105.0121 (2011). arXiv: 1105.0121. URL: http://arxiv.org/abs/1105.0121.

[43] Martin Ester et al. "A density-based algorithm for discovering clusters in large spatial databases with noise." In: *Kdd*. Vol. 96. 34. 1996, pp. 226–231.

[44] Peter J Rousseeuw. "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis". In: *Journal of computational and applied mathematics* 20 (1987), pp. 53–65.

[45] Chunhui Yuan and Haitao Yang. "Research on K-Value Selection Method of K-Means Clustering Algorithm". In: *J* 2.2 (2019), pp. 226–235.

[46] Svante Wold, Kim Esbensen, and Paul Geladi. "Principal component analysis". In: *Chemometrics and intelligent laboratory systems* 2.1-3 (1987), pp. 37–52.

[47] Laurens van der Maaten and Geoffrey Hinton. “Visualizing data using t-SNE”. In: *Journal of machine learning research* 9.Nov (2008), pp. 2579–2605.

[48] *Ecodriving, The Smart Driving Style*. URL: https://www.eltis.org/sites/default/files/Eco_Drive_6.pdf (visited on 10/25/2019).

[49] Marc Weber. *Automotive OBD-II Dataset*. 2019. DOI: 10.5445/IR/1000085073.

[50] *OBD Link MX+*. URL: http://www.obdlink.com/mxp/ (visited on 11/02/2019).

[51] *PLX Devices*. URL: https://www.plxdevices.com/Kiwi-4-OBD-Car-to-Smartphone-Connection-p/897346002092.htm (visited on 11/02/2019).

[52] *Beware Default Random Forest Importances*. URL: https://explained.ai/rf-importance/ (visited on 11/03/2019).

[53] *How to tune hyperparameters in tSNE*. URL: https://towardsdatascience.com/how-to-tune-hyperparameters-of-tsne-7c0596a18868 (visited on 09/19/2019).

[54] *Weather Underground*. URL: https://www.wunderground.com (visited on 10/23/2019).

[55] *Weather - In which ways does weather affect road safety?* URL: https://www.swov.nl/en/facts-figures/fact/weather-which-ways-does-weather-affect-road-safety (visited on 11/07/2019).

Appendix A

Descriptive Statistics

A.1 Descriptive statistics for DS_1

The Table A.1 shows the mean, standard deviation, minimum and maximum of the features derived. These statistics are useful for determining the limits for the scoring algorithm. The trips in this data set are driven in the south-west region of Germany.

Feature	**Mean**	**Std. Dev.**	**Min.**	**Max.**
Harsh Accel. per 100 km	5.9	5.5	0	36.49
Harsh Decel. per 100 km	7.07	8.51	0	58
PKE	3447	966	2148	8181
Engine Braking (%)	11	1.9	6.5	15.45
Cruise (%)	37	10	9.51	61.16
Idling (%)	7.02	3.84	0.52	21.63
High RPM (%)	1.71	2.50	0	10.11
Avg. Shift-up RPM	1989	73	1806	2182
% of High RPM up-shift	44	10	19	65

TABLE A.1: Descriptive statistics of the features derived for DS_1. (N = 63)

A.2 Descriptive statistics for DS_2

The Table A.2 shows the mean, standard deviation, minimum and maximum of the features derived for the second data set. These statistics were used to decide the limits for the scoring algorithm. The trips in this data set are driven in and around the city of Orlando by two drivers. Clustering was performed on this data set.

Feature	**Mean**	**Std. Dev.**	**Min.**	**Max.**
Harsh Accel. per 100 km	9	16	0	86
Harsh Decel. per 100 km	7.97	11.03	0	52.9
PKE	2446	1782	873	7409
Engine Braking (%)	9.5	7.7	2.95	36.5
Cruise (%)	60.7	20.7	11.18	86.26
Idling (%)	8.4	5.5	0	23.25
High RPM (%)	1.14	3.12	0	16.43
Avg. Shift-up RPM	2189	294	1739	3070
% of High RPM up-shift	20	16.7	0	71.43

TABLE A.2: Descriptive statistics for DS_2. (N = 35)

A.3 Descriptive statistics for DS_3

The Table A.3 shows the mean, standard deviation, minimum and maximum of the features derived for the second data set. This data set was collected to analyze the behaviour of the driver using the scoring algorithm. It is observed that engine braking is not registered. The signals for throttle position and instantaneous fuel consumption didn't work as expected. A total of 7 trips were collected around Enschede.

Feature	**Mean**	**Std. Dev.**	**Min.**	**Max.**
Harsh Accel. per 100 km	0	0	0	0
Harsh Decel. per 100 km	1.48	2.86	0	7.46
PKE	3288	426	2764	3946
Engine Braking (%)	0	0	0	0
Cruise (%)	26.35	11.53	0	37.33
Idling (%)	5.5	4.82	2.02	16.89
High RPM (%)	0	0	0	0
Avg. Shift-up RPM	1870	104	1704	2051
% of High RPM up-shift	0.7	2.11	0	5.97

TABLE A.3: Descriptive statistics for DS_3. (N = 7)

Appendix B

Correlation Maps

The Figure B.1 represents the correlation between the derived features and the fuel consumption for the first data set DS_1.

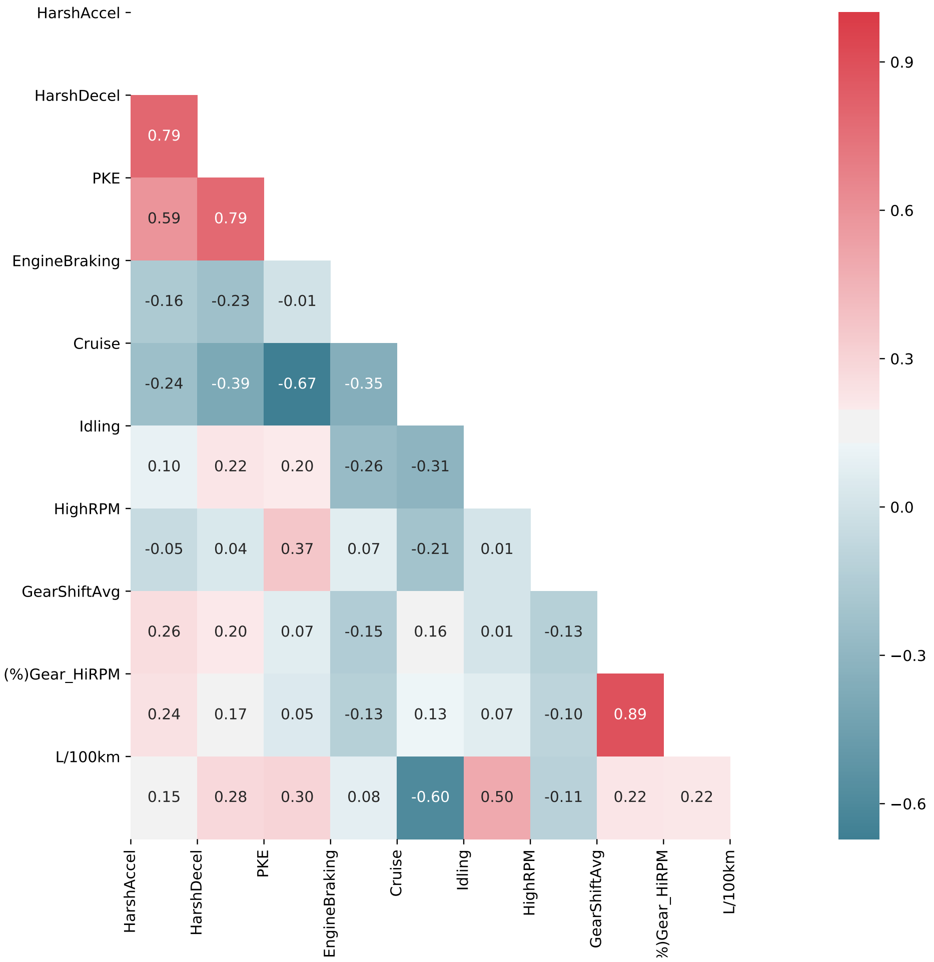


FIGURE B.1: Correlation map for DS_1

The Figure B.2 represents the correlation between the derived features and the fuel consumption for the second data set DS_2.

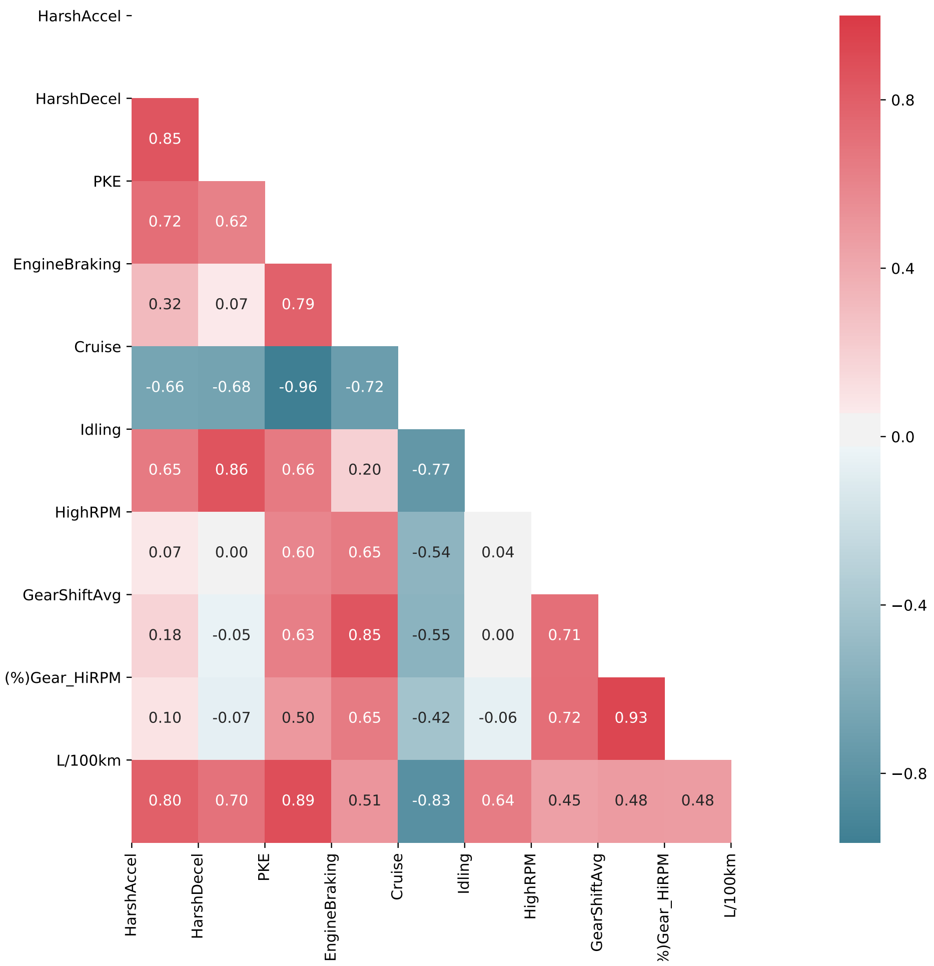


FIGURE B.2: Correlation map for DS_2

From both the Figures B.1 and B.2 it can be observed that engine braking which is supposed to have negative correlation with the fuel consumption, in fact has no correlation in DS_1 and has positive correlation in DS_2.