DRIVER DRIVING BEHAVIOUR PROFILING USING MACHINE LEARNING TECHNIQUES

LOW JIA MING

SESSION 2016/2017

FACULTY OF COMPUTING AND INFORMATICS
MULTIMEDIA UNIVERSITY

OCTOBER 2016

DRIVER DRIVING BEHAVIOUR PROFILING USING MACHINE LEARNING TECHNIQUES

BY

LOW JIA MING

SESSION 2016/2017

THIS PROJECT REPORT IS PREPARED FOR

FACULTY OF COMPUTING AND INFORMATICS
MULTIMEDIA UNIVERSITY
IN PARTIAL FULFILLMENT

FOR

BACHELOR OF COMPUTER SCIENCE

B.C.S (HONS) SOFTWARE ENGINEERING

FACULTY OF COMPUTING AND INFORMATICS

MULTIMEDIA UNIVERSITY

October 2016

DECLARATION

I hereby declare that the work has been done by myself and no portion of the work contained in this Thesis has been submitted in support of any application for any other degree or qualification on this or any other university or institution of learning.

_____

Low Jia Ming
Faculty of Computing and Informatics
Multimedia University
Date: 13:09:2016

ACKNOWLEDGEMENTS

I would like to say thank you to my supervisor Dr. Poo Kuan Hoong for guiding me throughout this Final Year Project. I also appreciate the advice and comments given by my co-supervisor, Dr. Ian Tan. The most important thing is the mental and physical support of my family members and friends. Lastly, I felt happy to have done this project. It is a great chance for me to learn new knowledge on machine learning. It is also my pleasure to have this chance to cooperate with my supervisor.

To my parents, my supervisors, and my friends.

ABSTRACT

This project proposed a driver driving behaviour profiling method based on the vehicle telemetric data and K-Means Algorithm. The proposed method collects the vehicle operational data and Global Positioning System (GPS) data. The vehicle telemetric data includes the vehicle speed, throttle position, engine speed, latitude coordinate, longitude coordinate, GPS speed, and others. The proposed method makes use of the On Board Diagnostic (OBD-II) interface, smartphone GPS device and Torque Lite version Android application to do the data collection. A feature will be added on to the dataset. The value of the feature determines whether current vehicle speed exceeded the speed limit of the road or not. K-Means algorithm will group the records into three clusters and each record will be labelled as good, medium or bad condition data according to the cluster. Based on the number of each condition data the driver had, the driver can be profiled as low risk, medium risk or high risk driver.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Basic Introduction

According to the statistics on drivers in Malaysia taken from Official Portal of Road Transport Department Malaysia shown in Figure1.1, the amount of drivers increases every year in Malaysia. The drivers got the license through driving examinations, but it does not mean that the drivers have good driving behaviour.

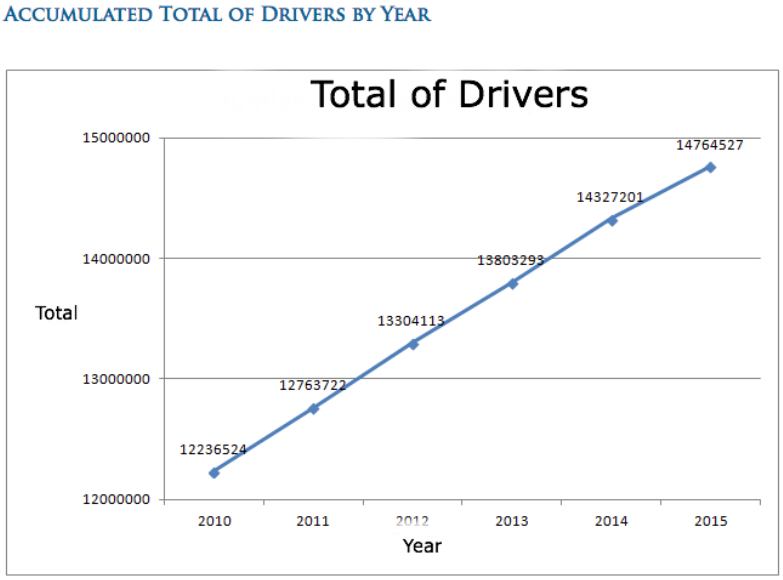


Figure 1.1: Accumulated total of drivers by year (source from: http://www.jpj.gov.my/)

According to the general road accident data in Malaysia taken from Malaysian Institute of Road Safety Research (MIROS) official website shown in Figure 1.2, Malaysia government put effort on reducing the amount of traffic incidents by introducing the new traffic laws and speed tracking system. However, the number of cases of road deaths does not drop significantly. It means that drivers' personal factors also related to the occurrence of traffic incident.

General Road Accident Data in Malaysia (1997 – 2014)

Year	Registered Vehicles	Population	Road Crashes	Road Deaths	Serious Injury	Slight Injury	Index per 10,000 Vehicles	Index per 100,000 Population	Indeks per billion VKT
1997	8,550,469	21,665,600	215,632	6,302	14,105	36,167	7.37	29.1	33.57
1998	9,141,357	22,179,500	211,037	5,740	12,068	37,896	6.28	25.8	28.75
1999	9,929,951	22,711,900	223,166	5,794	10,366	36,777	5.83	25.5	26.79
2000	10,598,804	23,263,600	250,429	6,035	9,790	34,375	5.69	26.0	26.25
2001	11,302,545	23,795,300	265,175	5,849	8,680	35,944	5.17	25.1	23.93
2002	12,068,144	24,526,500	279,711	5,891	8,425	35,236	4.9	25.3	22.71
2003	12,819,248	25,048,300	298,653	6,286	9,040	37,415	4.9	25.1	22.77
2004	13,828,889	25,580,000	326,815	6,228	9,218	38,645	4.52	24.3	21.1
2005	15,026,660	26,130,000	328,264	6,200	9,395	31,417	4.18	23.7	19.58
2006	15,790,732	26,640,000	341,252	6,287	9,253	19,885	3.98	23.6	18.69
2007	16,813,943	27,170,000	363,319	6,282	9,273	18,444	3.74	23.1	17.6
2008	17,971,901	27,730,000	373,071	6,527	8,868	16,879	3.63	23.5	17.65
2009	19,016,782	28,310,000	397,330	6,745	8,849	15,823	3.55	23.8	17.27
2010	20,188,565	28,910,000	414,421	6,872	7,781	13,616	3.4	23.8	16.21
2011	21,401,269	29,000,000	449,040	6,877	6,328	12,365	3.21	23.7	14.68
2012	22,702,221	29,300,000	462,423	6,917	5,868	11,654	3.05	23.6	13.35
2013	23,819,256	29,947,600	477,204	6,915	4,597	8,388	2.90	23.1	12.19
2014	25,101,192	30,300,000	476,196	6,674	4,432	8,598	2.66	22.0	10.64

Figure 1.2: General Road Accident Data in Malaysia (1997 - 2014) (source from: https://www.miros.gov.my/)

The driver characteristics and the occurrence of traffic incident is interrelated. To further reduce the number of accidents, the safety equipment of the vehicle needs to be improved as well as the road regulations, but also pay attention to driver behaviour. However, the behaviour of the driver is hard to be identified. The driver behaviour is affected by environment, vehicle condition and the mental or physical state.(Miyaji, Danno, & Oguri, 2008) One of the ways to identify driver behaviour is using the vehicle telemetric data.

1.2 Research Motivation

This project might help insurance company to de-tariff the Motor Insurance in Malaysia. However, the insurance companies cannot obtain the vehicle telemetric data from customers directly. They cannot force their customers to drive without knowing that they are monitoring the customer's vehicle telemetric data.

Currently, new high end cars do send driving data back to the manufacturers. Consumer group Federation Internationale de I'Automobile (FIA) exposed Bayerische Motoren Werke AG (BMW) received data from their manufactured cars. The data actually can analyse driver driving behaviour.(Stupp, 2015) So, the model built in this project can possibly be used when these insurance companies want to classify drivers. However, the insurance companies may classify the drivers with lesser data.

1.3 Project Objective

1. To identify the features that contributes to the accuracy of the classification of the driver behaviour analysis from the vehicle telemetric data.
2. To classify each vehicle telemetric records.
3. To profile the drivers based on the labelled vehicle telemetric data.

1.4 Project Scope

This project focuses on the driver driving behaviour profiling. Actual vehicle telemetric data are captured by sensor. The vehicle telemetric data will be collected and preprocessed before analysing. This project requires a vehicle that supports the On Board Diagnostic (OBD-II) device. Each driver is required to drive the vehicle for at least 8 minutes to collect data. The data will be recorded every second. For each driver, there are at least 480 records in the dataset. K-Means algorithm is implemented in this project to cluster the vehicle operation records. Each record will be labelled as good, medium or bad condition. Based on the labelled records, the drivers will be categorized to three classes. The classes are low risk, medium risk and high risk.

1.5 Project Plan

Proper time and resource management is important to complete this project. Table 1.5 shows a brief description of the project timeline.

Task \Week	2	3	4	5	6	7	8	9	10	11	12
Data Acquisition	■	■	■	■	■						
Literature Review		■	■	■	■						
Data Analysis using KNIME						■	■	■			
Driver Driving Behaviour Profiling								■	■		
Analysing Experimental Result									■	■	
Documentation and Report									■	■	■
Report Submission											■

Table 1.1: Project timeline of Final Year Project I for Trimester 1 2016/2017

CHAPTER 2

RELATED WORK

2.1 Introduction

Various literatures have introduced driving behaviour analysis. Driving behaviour can be influenced by the emotion of the drivers. The driver may drive faster when they feel angry or sad. The distracted drivers cause most of the accidents. The drivers may be distracted by phone call while driving. Driver behaviour is hard to be determined by the speech emotion of the driver. (Kamaruddin & Wahab, 2010)

The driving environment will also affect the driver behaviour. Some of the drivers will pass through the intersection without stopping and observing the surrounding condition. The decision of drivers for passing through certain road condition will reflect the drivers' driving behaviour.

The On Board Diagnostic (OBD) information will directly reflect the driving behaviour of the driver. Vehicle speed can determine the current driving state whether accelerating or remaining safe. Engine speed can determine the effciency of the vehicle operation. The number of fuel sent to the engine is determined by the throttle position. Suitable throttle position will ensure that the engine operates efficiently. Inappropriate throttle position will cause incomplete combustion of fuel and air pollution.(Chen, Pan, & Lu, 2015)

2.2 Driver Behaviour Analysis through Speech Emotional Understanding

This paper was proposed by Kamaruddin and Wahab (2010). The researchers analysed the driver behaviour state (DBS) based on the emotion of the driver when the driver was driving. The emotion of the driver can be detected through speech. The researches used the Berlin dataset and NAW dataset as training set. The Berlin dataset

and NAW dataset are standard dataset for speech emotion recognition and have been used by many researchers. The proposed method used the Generic Self-organizing Fuzzy Neural Network (GensoFNN) as a classifier for identification purpose.

2.2.1 Driver Behaviour Analysis Method

This proposed method collected data from 11 adults (3 female and 8 male). The drivers have at least two years of driving experience and drive at least 10 hours per week. The drivers are required to do four designed actions. The actions are:

1. Driver talking through the mobile phone while driving.
2. Driver feeling sleepy.
3. Driver laugh while driving
4. Driver in the initial driving exercise where the driver is in neural state of emotion.

A microphone is embedded in the vehicle to collect the speech of the driver while driving. The experiments to analyse the three DBS of talking, laughing, and sleepy were conducted. The researchers use the Berlin dataset and NAW dataset to relate the three DBS with the angry, happy, and sad emotion.

2.2.2 Experimental Result

The result shows that the sleepy DBS can be recognised as sad emotion consistently. However, the talking and laughing DBS gave mixed results. It means that more work need to be conducted for better classification of the two DBS. The accuracy of sleepy emotion detection is up to 65% using the proposed speech emotion recognition system.

2.3 Driver Behaviour Analysis and Route Recognition by Hidden Markov Models

This paper was proposed by Sathyanarayana, Boyraz, and Hansen (2008). This paper introduced the driver behaviour modelling using Hidden Markov Models (HMM)

in Bottom-to-Top Approach and Top-to-Bottom Approach.

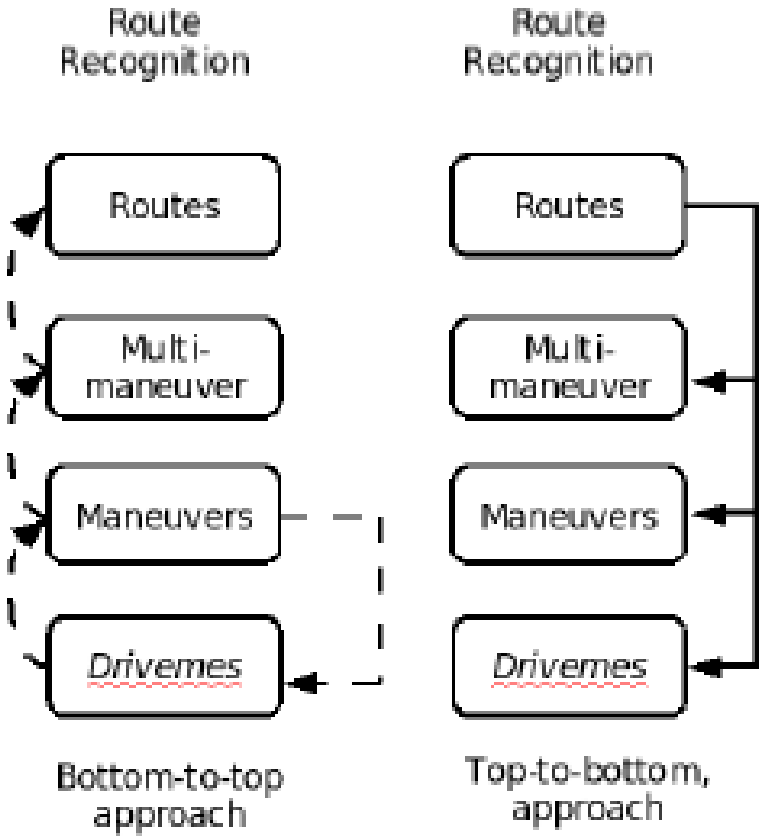


Figure 2.1: Hierarchy among the units of route recognition (Sathyanarayana et al., 2008)

In the Bottom-to-top approach, the manoeuvres are the smallest unit to be recognised by the algorithms. The driver behaviour can be discovered from the recognised manoeuvres and be used to build the manoeuvre models. The multi-manoeuvre models use the manoeuvres model to be built and finally become a complete route. The Bottom-to-Top approach is to detect the distraction of driver based on the comparison of the neutral state vehicle telemetries in the known manoeuvres.

In Top-to-Bottom approach, it is an opposite approach in driving behaviour modelling. A single HMM is used to parse the route to the individual meaningful parts like manoeuvres and states. Three main clusters will be identified to represent the driver driving behaviour in intersection. The three main clusters are determined

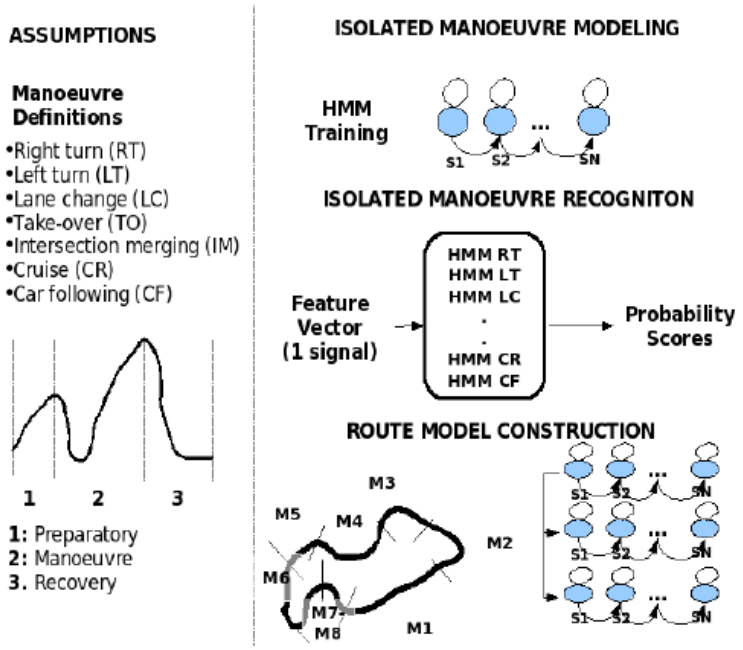


Figure 2.2: Bottom-to-top approach for route model construction (Sathyanarayana et al., 2008)

by the manoeuvres. The manoeuvres are lane change, left turn, and right turn. The Top-to-Bottom approach is to recognise the route based on vehicle information data.

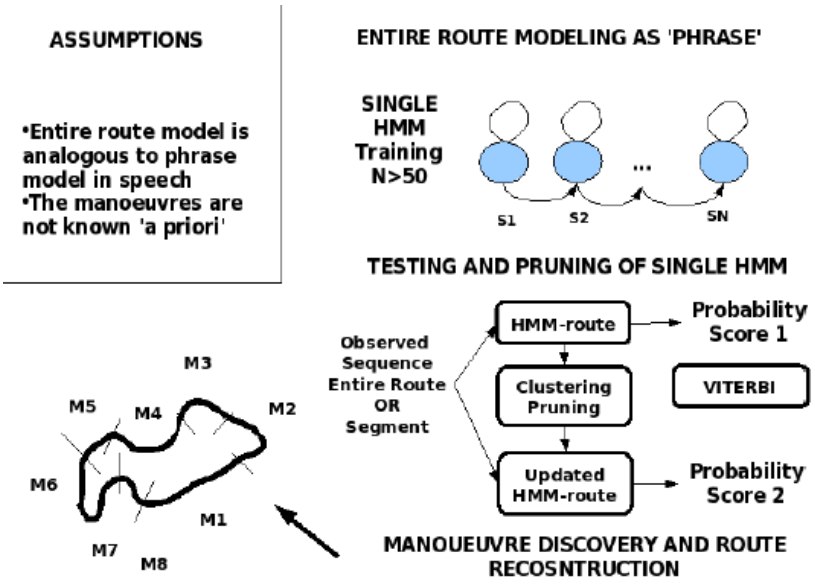


Figure 2.3: Top-to-bottom approach for route model construction (Sathyanarayana et al., 2008)

2.3.1 Data Collection

The proposed method used UTDrive Vehicle to collect the data. The UTDrive Vehicle is converted from a Toyota RAV4 vehicle. The UTDrive Vehicle is equipped camera to capture the driver and the road. The driver speech also been recorded by the equipped microphone in the vehicle. Distance sensors using laser and GPS for position measurement are also equipped. The CAN-Bus is used to collect the vehicle speed, steering wheel angle, and brake/gas information. The vehicle also equipped gas/brake pedal pressure sensors.

The experiment conducted in two different areas in order to collect two different scenarios data. The residential and commercial area including the right turn, left turn and lane change were selected for drivers to driver through. The drivers were required to drive two times with neutral driving and distracted driving.

2.3.2 Experimental Result

Three different types of manoeuvres (left turn, right turn, and lane change) were recognised by using the three CAN-Bus signals (vehicle speed, steering angle, and brake force). On the other hand, the distraction detection was recognised and having 95% accuracy.

2.4 Driving Behaviour Analysis Based on Vehicle OBD Information and AdaBoost Algorithms

This paper was proposed by Shi-Huang et al.(2015). The researchers analysed the drivers' behaviour based on the On Board Diagnostic (OBD) information and using the AdaBoost Algorithms to create the driving behaviour classification. Finally, the experimental results show the correctness of the proposed driving behaviour analysis method has 99.8% accuracy rate in various driving simulations.

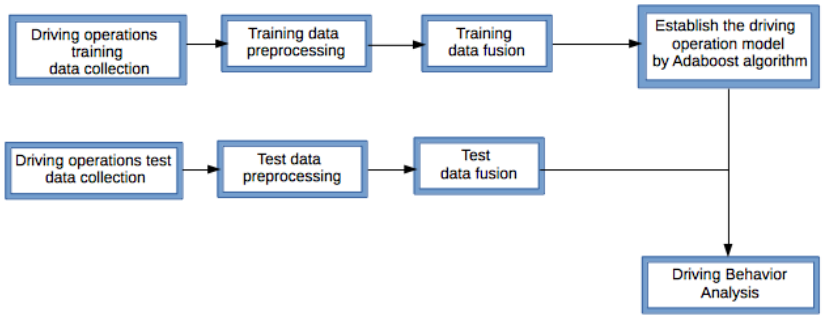


Figure 2.4: The flowchart of the proposed driving behaviour analysis method.

2.4.1 AdaBoost Algorithms

AdaBoost is a classification machine learning algorithm. The AdaBoost algorithms are used for forming a strong classifier by combining large number of weak classifier. There are three different types of AdaBoost algorithms. They are Gentle AdaBoost, Modest AdaBoost and Real AdaBoost.

2.4.2 Driving Behaviour Analysis Method

This proposed method used the OBD-II system in the vehicle and EZ-SCAN5 as the OBD-II to Bluetooth adapter to collect OBD-II vehicle information data. OBD-II is proposed in 1996 to replace the OBD-I system. The OBD-II system is implemented in every vehicle under the Environmental Protection Agency (EPA) regulation in USA since 1996. When the air-pollution contents exhausted by the vehicle exceed the minimum level, the OBD-II system of the particular vehicle will generate the Diagnostic Trouble Code (DTC) message and Check Engine light will display on the vehicle dashboard. In order to communicate with the OBD-II system of the vehicle, EZ-SCAN5 as a OBD-II to Bluetooth adapter is required. The EZ-SCAN5 support most of the OBD-II communication protocols, such as SAE J1850 PWM, SAE J1850 VPW, ISO 9141-2, ISO 14230-4 KWP, and ISO 15765-4 CAN. If the adapter does not support the communication protocol of the OBD-II system, the OBD-II vehicle information will not be able to retrieve.

This proposed method collected vehicle speed, engine speed (RPM), throttle position and engine load as the OBD-II vehicle information data. According to the engine characteristic curve, the proposed method developed two criteria for the data collection.

1. The normal vehicle condition data

   The relative ratio of the vehicle speed and the engine speed is remained in a range that between 0.9 and 1.3. The result was tested in the same gear. The relative ratio of the engine speed and throttle valve is remained in a range that between 0.9 and 1.3. The engine load is remained between 20% and 50%.

2. The bad vehicle condition data

   The relative ratio of the vehicle speed and the engine speed is out of the range that between 0.9 and 1.3. The result was tested in the same gear. The relative ratio of the engine speed and throttle valve is out of the range that between 0.9 and 1.3. The engine load is out of the range that between 20% and 50%.

2.4.3 Vehicle OBD information Data Preprocessing

The proposed method used three characteristics. The three characteristics are the relative ratio of the vehicle speed and engine speed, the relative ratio of throttle position and engine speed, and engine load. Using the characteristics to analyse the current state of the driving behaviour whether the driver is in safe state or dangerous state. The proposed method needed to compute the change rate of vehicle speed, engine speed and throttle position in the first step. The calculation is shown in Equation(2.1), where $t_2 - t_1 = 1$.

$$D(t) = \frac{data(t_2) - data(t_1)}{t_2 - t_1} \tag{2.1}$$

The next step is to calculate the relative ratio of the vehicle speed and engine speed, and the relative ratio of the throttle position and engine speed. The calculation is shown in the Equation(2.2) and (2.3).

$$R_{cz}(t) = \frac{cs(t)}{220} \div \frac{zs(t)}{8000} \tag{2.2}$$

where $R_{cz}(t)$ is the relative ratio of engine speed and vehicle speed. $cs(t)$ is the vehicle speed at time t. The 220 is the value of maximum vehicle speed. $zs(t)$ is the engine speed a time t. The 8000 is the value of maximum engine speed of the vehicle.

$$R_{jz}(t) = \frac{jq'(t)}{max(jq')} \div \frac{zs'(t)}{max(zs'(t))} \tag{2.3}$$

where $R_{jz}(t)$ is the relative ratio of engine speed and throttle position. The $jq'(t)$ is the change rate of the throttle position at time t. $zs'(t)$ is the change rate of the engine speed at time t. The $max(jq'(t))$ and $max(zs'(t))$ is the maximum value of the change rate of throttle position and engine speed, respectively.

Based on the two computed features and engine load, these three features were combined to determine the vehicle data is normal or bad driving behaviour.

2.4.4 Experimental Result

The proposed method used the toolkit-GML-AdaBoost-matlab of MATLAB to execute data preprocessing and driving behaviour modelling. The preprocessed data is tested on three different types of the AdaBoost algorithms. They are Gentle AdaBoost, Modest AdaBoost and Real AdaBoost. The Real AdaBoost is better than other AdaBoost algorithms with the highest accuracy rate, 99.8%.

CHAPTER 3

REQUIREMENT

3.1 Hardware Requirement Introduction

In this project, a smartphone is required. The smartphone must have GPS and Bluetooth device. An OBD-II to Bluetooth adapter, ELM327 is required to use in data collection. The model of vehicle used to collect driver data should be year 2006 onward. Some of the brand cannot be supported by the ELM327 due to the protocols mismatch. The ELM327 supports most of the vehicle model with year 2008 in Proton brand, Toyota brand, and Honda brand. However, the latest model of the vehicle may not be supported due to the protocols mismatch problem also.

3.1.1 Vehicle OBD system

The OBD system is also called OBD-II, was proposed in 1996. In 1996, all the cars manufactured in United State (US) were required to equip OBD-II and the cars without OBD-II prohibited selling in US. The purpose to have OBD-II specifications is to diagnose engine problem. Environment Protection Agency (EPA) and the state of California used the specification to meet the emission standards. Since 1996, all the cars in US are required to be equipped with OBD-II to establish the EPA regulation.

The usage of the OBD-II is important for detecting the vehicle exhaustion. If the vehicle is exhaust high level of air-pollution content, Diagnostic Trouble Codes (DTCs) will be generated by the OBD-II and a Check Engine Light will be displayed on vehicle dashboard. OBD-II will store these DTCs into the Engine Control Unit memory. An OBD-II scanning tool can access the ECU to retrieve the DTCs.

The OBD-II is usually installed under the vehicle dashboard and above the pedals. Figure 3.1 shows the OBD-II socket of Toyota Vios year 2007 model.

Figure 3.1: OBD-II socket of the Toyota Vios year 2007 model.

3.1.2 ELM327

Super Mini ELM327 Bluetooth OBD-II is an OBD-II scanning tool produced by ELM Electronics. It is a programmed micro-controller to communicate with the OBD-II port of the vehicle. The ELM327 supports most of OBD-II protocols. ELM327 also contains the Bluetooth adapter. The ELM327 needs to be plugged to the OBD-II port.

3.2 Software Requirement Introduction

In order to collect the vehicle telemetric data, Torque(Lite), an Android application needs to be installed in the smartphone. Vehicle telemetric data will be stored in smartphone with Comma Separated Values (CSV) file format. Google Drive will be utilized to store the CSV file for backup purpose.

In order to add on the new features to the dataset, Google Sheet and LibreOffice are used for modifying the dataset.

Google Earth are used for verify the road structure that drivers drove through.

Figure 3.2: Super Mini ELM327 Bluetooth OBD-II.

The Torque(Lite) can convert the CSV file to Keyhole Markup Language (KML) file.

LibreOffice are applied in this project to do the final modification of the dataset. After the modification, the dataset will be stored as CSV format file. KNIME Analytics Platform is used for clustering the collected vehicle telemetric data by performing K-Means Algorithm. The result will be visualized by using Tableau.

3.2.1 Torque(Lite)

Torque Lite version is a free android application and it can be installed in smartphone from Google Play Store. The application will communicate with the ELM327 through the Bluetooth connection. The application will collect the data received from the ELM327 and save the data into a CSV file in the smartphone.

After vehicle ignition is on, ELM327 need to connect with the smartphone that operating Torque via Bluetooth. Once the smartphone connected with the adapter, Torque will choose the protocol that matches with the OBD-II system. After the matching successes, Torque will start to read the sensor information from the ECU. In order to collect the GPS data, smartphone's location service is required to be enabled at the same time. When the smartphone received GPS signal and connected with the adapter, there are four flashing icons at the top right of the main screen will stay solid with blue color. Figure 3.3 shows a screen-shot of the Torque(Lite).

Figure 3.3: Screen-shot of the Torque(Lite).

3.2.2 KNIME Analytics Platform

KNIME stands for KoNstanz Information MinEr. KNIME Analytics Platform is a product of KNIME. KNIME Analytics Platform is an open platform for data analysis. It is a perfect tool for data scientists. The version of KNIME Analytics Platform used in this project is version 3.2.0.

KNIME Analytics Platform contains 1000 modules, hundreds of ready-to-run examples, a set of integrated tools, and a list of advance algorithms available. User can build a machine learning experiment by dragging and dropping the related modules into the project. User needs to link and configure the module before execution.

This project uses the module of the KNIME Analytics Platform to complete the data preprocessing and driving behaviour modelling. K-Means algorithm module is used in this project to classify the driving behaviour.(Berthold et al., 2007)

3.2.3 K-Means Algorithm

K-Means algorithm is a popular clustering algorithm. The algorithm is used for automatically grouping data into cohesive 'clusters'. Few points will be initialized randomly. The number of clusters to be generated determines the amount of initial points. For example, in order to obtain three clusters as output, three points is required to be initialized in the first step. The points are called cluster centroids (Ng, 2016).

The second step of the algorithm is an inner loop for performing two functions. The functions are shown below:

1. Assign each training example to the closest cluster.
2. Move the cluster centroids to the mean of the points assigned to it.

This looping will be ended when the new iterations do not make the cluster centroids moved.

CHAPTER 4

DESIGN

4.1 Introduction

In this project, the driving behaviour analysis method contains driving operation data acquisition module, data preprocessing module, data fusion module and K-Means algorithm module. The raw data will be preprocessed. Each driver's vehicle telemetric data will be concatenated into a single CSV file. The new CSV file will be categorized to three groups by using K-Means algorithm. Figure 4.1 shows the flow of entire project.

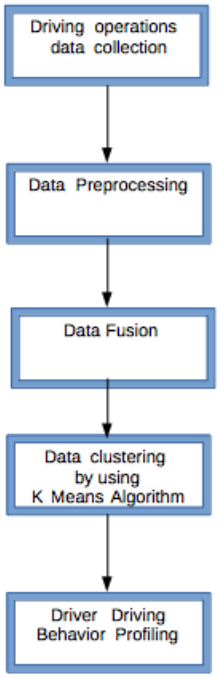


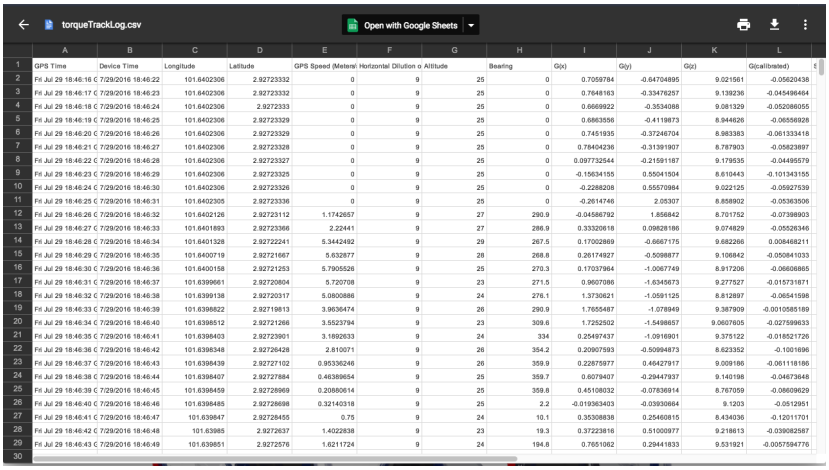
Figure 4.1: The flow of the driver driving behaviour profiling method.

4.2 Data Acquisition

In this project, few drivers have been invited to participate in the data collection. The ELM327 device needs to be connected with the vehicle OBD-II socket. The

smartphone with the Torque(Lite) needs to be put in the vehicle. It needs to connect with the ELM327 device and GPS location service. Torque(Lite) logging function needs to be triggered and stopped in each driver driving session.

Vehicle telemetric data will be recorded in every second. The drivers are requested to drive at least 8 minutes. At least 480 vehicle operation records can be collected from each driver. All drivers were driving in different path. The raw data will be uploaded and stored in Google Drive. Figure 4.2 shows the view of the raw data in Google Drive.



	A	B	C	D	E	F	G	H	I	J	K	L
1	GPS Time	Device Time	Longitude	Latitude	GPS Speed (Meters/	Horizontal Dilution o	Altitude	Bearing	G(x)	G(y)	G(z)	G(calibrated)
2	Fri Jul 29 18:46:16 C	7/29/2016 18:46:22	101.6402306	2.9272332	0	9	25	0	0.7059784	-0.64704895	9.021561	-0.05620438
3	Fri Jul 29 18:46:17 C	7/29/2016 18:46:23	101.6402306	2.9272332	0	9	25	0	0.7648163	-0.33476257	9.139236	-0.045496464
4	Fri Jul 29 18:46:18 C	7/29/2016 18:46:24	101.6402306	2.9272333	0	9	25	0	0.6669922	-0.3534088	9.081329	-0.052086055
5	Fri Jul 29 18:46:19 C	7/29/2016 18:46:25	101.6402306	2.9272329	0	9	25	0	0.6863556	-0.4119873	8.944626	-0.06556928
6	Fri Jul 29 18:46:20 C	7/29/2016 18:46:26	101.6402306	2.9272329	0	9	25	0	0.7451935	-0.37246704	8.983383	-0.061333418
7	Fri Jul 29 18:46:21 C	7/29/2016 18:46:27	101.6402306	2.9272328	0	9	25	0	0.78404236	-0.31391907	8.787903	-0.05823897
8	Fri Jul 29 18:46:22 C	7/29/2016 18:46:28	101.6402306	2.9272327	0	9	25	0	0.097732544	-0.21591187	9.179535	-0.04495579
9	Fri Jul 29 18:46:23 C	7/29/2016 18:46:29	101.6402306	2.9272325	0	9	25	0	-0.15634155	0.55041504	8.610443	-0.101343155
10	Fri Jul 29 18:46:24 C	7/29/2016 18:46:30	101.6402306	2.9272326	0	9	25	0	-0.2288208	0.55570984	9.022125	-0.05927539
11	Fri Jul 29 18:46:25 C	7/29/2016 18:46:31	101.6402305	2.9272336	0	9	25	0	-0.2614746	2.05307	8.858902	-0.05363506
12	Fri Jul 29 18:46:26 C	7/29/2016 18:46:32	101.6402126	2.9272112	1.1742657	9	27	290.9	-0.04586792	1.856842	8.701752	-0.07398903
13	Fri Jul 29 18:46:27 C	7/29/2016 18:46:33	101.6401893	2.9272366	2.22441	9	27	286.9	0.33320618	0.09828186	9.074829	-0.05526346
14	Fri Jul 29 18:46:28 C	7/29/2016 18:46:34	101.6401328	2.9272224	5.3442492	9	29	267.5	0.17002869	-0.6667175	9.682266	0.008468211
15	Fri Jul 29 18:46:29 C	7/29/2016 18:46:35	101.6400719	2.9272167	5.632877	9	28	268.8	0.26174927	-0.5098877	9.106842	-0.050841033
16	Fri Jul 29 18:46:30 C	7/29/2016 18:46:36	101.6400158	2.9272125	5.7905526	9	25	270.3	0.17037964	-1.0067749	8.917206	-0.06608665
17	Fri Jul 29 18:46:31 C	7/29/2016 18:46:37	101.6399661	2.9272084	5.720708	9	23	271.5	0.9607086	-1.6345673	9.277527	-0.015731871
18	Fri Jul 29 18:46:32 C	7/29/2016 18:46:38	101.6399138	2.9272031	5.0800886	9	24	276.1	1.3730621	-1.0591125	8.812897	-0.06541598
19	Fri Jul 29 18:46:33 C	7/29/2016 18:46:39	101.6398622	2.9271981	3.9636474	9	26	290.9	1.7655487	-1.078949	9.387909	-0.0010585189
20	Fri Jul 29 18:46:34 C	7/29/2016 18:46:40	101.6398512	2.9272126	3.5523794	9	23	309.6	1.7252502	-1.5498657	9.0607605	-0.027599633
21	Fri Jul 29 18:46:35 C	7/29/2016 18:46:41	101.6398403	2.9272301	3.1892633	9	24	334	0.25497437	-1.0916901	9.375122	-0.018521726
22	Fri Jul 29 18:46:36 C	7/29/2016 18:46:42	101.6398348	2.9272642	2.810071	9	26	354.2	0.20907593	-0.50994873	8.623352	-0.1001696
23	Fri Jul 29 18:46:37 C	7/29/2016 18:46:43	101.6398439	2.9272702	0.95336246	9	26	359.9	0.22875977	0.46427917	9.009186	-0.061118186
24	Fri Jul 29 18:46:38 C	7/29/2016 18:46:44	101.6398407	2.9272884	0.46389654	9	25	359.7	0.6079407	-0.29447937	9.140198	-0.04673848
25	Fri Jul 29 18:46:39 C	7/29/2016 18:46:45	101.6398459	2.9272869	0.20860614	9	25	359.8	0.45108032	-0.07836914	8.767059	-0.08609629
26	Fri Jul 29 18:46:40 C	7/29/2016 18:46:46	101.6398485	2.9272698	0.32140318	9	25	2.2	-0.019363403	-0.03930664	9.1203	-0.0512951
27	Fri Jul 29 18:46:41 C	7/29/2016 18:46:47	101.639847	2.9272455	0.75	9	24	10.1	0.35308838	0.25460815	8.434036	-0.12011701
28	Fri Jul 29 18:46:42 C	7/29/2016 18:46:48	101.63985	2.9272637	1.4022838	9	23	19.3	0.37223816	0.51000977	9.218613	-0.039082587
29	Fri Jul 29 18:46:43 C	7/29/2016 18:46:49	101.639851	2.9272576	1.6211724	9	24	194.8	0.7651062	0.29441833	9.531921	-0.0057594776
30												

Figure 4.2: The raw data file is opened in Google Drive.

For this project, the vehicle OBD-II information data and GPS data will be collected and store in the same file. Some of the features are vehicle speed, engine speed, engine coolant temperature, throttle position, latitude coordinate, longitude coordinate, GPS speed, GPS altitude, horizontal dilution of precision, GPS bearing, and GPS acceleration. Those data will be written in a same file.

A new feature, speed test can be added into the data. Speed test is a value to determine whether the driver exceeded the speed limit or not at the particular time frame. The value of the speed test is '-1' or '1'. '-1' means the driver exceeded the speed limit at the particular time frame. '1' is on the contrast. The speed limit will be determined by the road condition. The road condition includes straight road, curve road, traffic light intersection, intersection, roundabout, and state road. The road

condition will be referred from the Google Earth. The KML file converted by the Torque(Lite) will show the path of the driver drove. In Figure 4.2, each record of the vehicle telemetric data will be displayed as a green arrow and marked on the map.

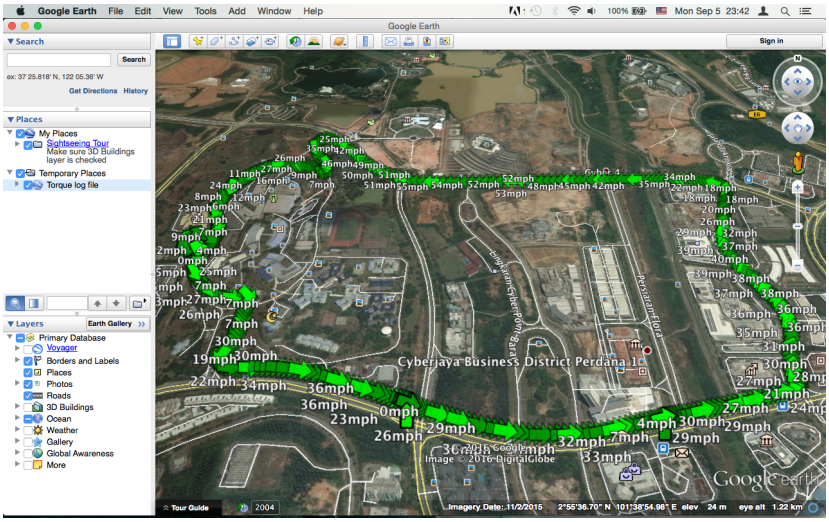


Figure 4.3: The Map View in Google Earth for the vehicle telemetric data.

In order to identify the intersection on straight road or differentiate the traffic light intersection and normal intersection, Google Earth provided the Street View function that allows observing the surroundings of the particular place. In Figure 4.4, one of the traffic light intersection examples displayed in Google Earth.

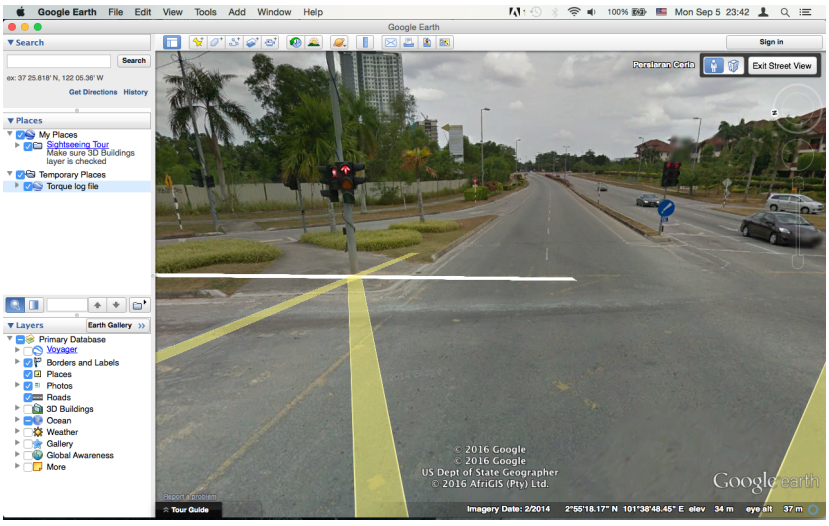


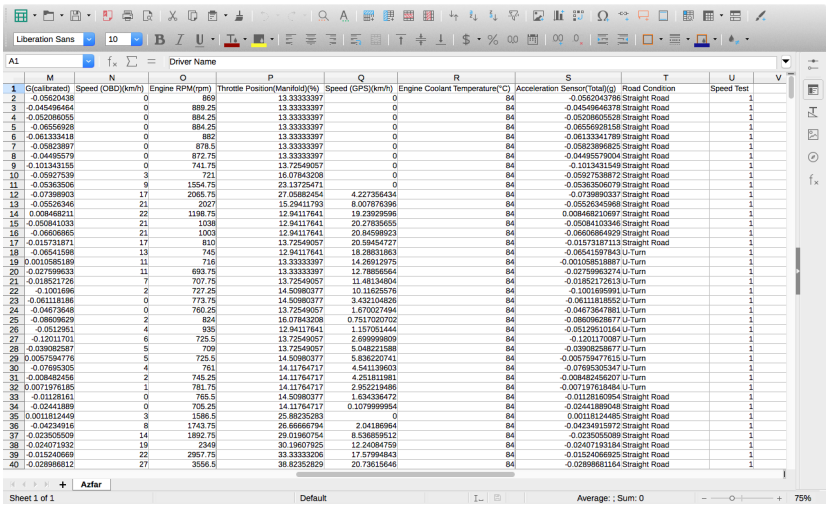
Figure 4.4: The Street View in Google Earth at the traffic light intersection.

In Malaysia, speed limit of expressways by default is 110 km/h, but it will be 90km/h in crosswind area, mountainous stretches area, and urban area. Speed limit

of state road or federal road is 80km/h, town area will be 60km/h. Due to some road condition, a 40km/h or 50km/h speed limit sign board will be placed in the area of having curve road.

Drivers suppose to stop at the intersection and observe the surroundings vehicle movement before turning to another route. So, speed limit of intersection area will be set as 30km/h in this project to determine the drivers' speed limit condition. The intersections having traffic light will be applied with 35km/h speed limit. Drivers should not drive through the traffic light intersection area with high speed. In some circumstances, drivers will accelerate when the traffic light changes to amber. As the result, the drivers will stop the car uncomfortably due to this insufficient time period. The drivers need to make decision at the stop-line either to pass through the intersection after red signals or brake hard in front of the intersection. This action will increase the potential of accident occurances.(Kulanthayan, Phang, & Hayaticoncatenate, 2007)

LibreOffice and Google Sheet are used to add on the speed test feature and the road condition feature. The driver name is also added to the vehicle telemetric data of the particular driver. Figure 4.5 shown the vehicle telemetric data after added on new features.



	M	N	O	P	Q	R	S	T	U	V
1	G(calibrated)	Speed (OBD)(km/h)	Engine RPM(rpm)	Throttle Position(Manifold)(%)	Speed (GPS)(km/h)	Engine Coolant Temperature(°C)	Acceleration Sensor(Total)(g)	Road Condition	Speed Test	
2	-0.05620438	0	869	13.33333397	0	84	-0.056204378	Straight Road	1	
3	-0.045496464	0	889.25	13.33333397	0	84	-0.04549646378	Straight Road	1	
4	-0.052086055	0	884.25	13.33333397	0	84	-0.05208605528	Straight Road	1	
5	-0.06556928	0	884.25	13.33333397	0	84	-0.06556928158	Straight Road	1	
6	-0.061333418	0	882	13.33333397	0	84	-0.06133341789	Straight Road	1	
7	-0.05823897	0	878.5	13.33333397	0	84	-0.05823896825	Straight Road	1	
8	-0.044855579	0	872.75	13.33333397	0	84	-0.04485579004	Straight Road	1	
9	-0.101343155	0	741.75	13.72549057	0	84	-0.1013431549	Straight Road	1	
10	-0.05927539	3	721	16.07843208	0	84	-0.05927538872	Straight Road	1	
11	-0.05363506	9	1554.75	23.13725471	0	84	-0.05363506079	Straight Road	1	
12	-0.07398903	17	2065.75	27.05882454	4.227356434	84	-0.0739890337	Straight Road	1	
13	-0.05526346	21	2027	15.29411793	8.007876396	84	-0.05526345968	Straight Road	1	
14	0.008468211	22	1198.75	12.94117641	19.23929596	84	0.008468210697	Straight Road	1	
15	-0.050841033	21	1038	12.94117641	20.27635655	84	-0.0508410334	Straight Road	1	
16	-0.06606865	21	1003	12.94117641	20.84598923	84	-0.06606864929	Straight Road	1	
17	-0.015731871	17	810	13.72549057	20.59454727	84	-0.01573187113	Straight Road	1	
18	-0.06541598	13	745	12.94117641	18.28831863	84	-0.0654159784	U-Turn	1	
19	0.0010585189	11	716	13.33333397	14.26912975	84	-0.001058518887	U-Turn	1	
20	-0.027599633	11	693.75	13.33333397	12.78856564	84	-0.0275996274	U-Turn	1	
21	-0.018521726	7	707.75	13.72549057	11.48134804	84	-0.0185217261	U-Turn	1	
22	-0.1001696	2	727.25	14.50980377	10.11625576	84	-0.1001695991	U-Turn	1	
23	-0.061118186	0	773.75	14.50980377	3.432104826	84	-0.0611181855	U-Turn	1	
24	-0.04673648	0	760.25	13.72549057	1.670027494	84	-0.0467364788	U-Turn	1	
25	-0.08609629	2	824	16.07843208	0.7517020702	84	-0.0860962867	U-Turn	1	
26	-0.0512951	4	935	12.94117641	1.157061444	84	-0.0512951016	U-Turn	1	
27	-0.12011701	6	725.5	13.72549057	2.699999609	84	-0.120117008	U-Turn	1	
28	-0.039082587	5	709	13.72549057	5.048221588	84	-0.0390825877	U-Turn	1	
29	0.0057594776	5	725.5	14.50980377	5.836220741	84	-0.005759477615	U-Turn	1	
30	-0.07695305	4	761	14.11764717	4.541139603	84	-0.0769530534	U-Turn	1	
31	-0.008482456	2	745.25	14.11764717	4.251811981	84	-0.0084824562	U-Turn	1	
32	0.0071976185	1	781.75	14.11764717	2.952219486	84	-0.007197618	U-Turn	1	
33	-0.01128161	0	765.5	14.50980377	1.634336472	84	-0.0112816095	Straight Road	1	
34	-0.02441889	0	705.25	14.11764717	0.1079999954	84	-0.0244188904	Straight Road	1	
35	0.0011812449	3	1586.5	25.88235283	0	84	0.001181244	Straight Road	1	
36	-0.04234916	8	1743.75	26.66666794	2.04186964	84	-0.0423491597	Straight Road	1	
37	-0.023505509	14	1892.75	29.01960754	8.536859512	84	-0.0235055089	Straight Road	1	
38	-0.024071932	19	2349	30.19607925	12.24084759	84	-0.0240719318	Straight Road	1	
39	-0.015240669	22	2957.75	33.33333206	17.57994843	84	-0.0152406925	Straight Road	1	
40	-0.028986812	27	3556.5	38.82352829	20.73615646	84	-0.0289868116	Straight Road	1	

Figure 4.5: The vehicle telemetric data is added on the speed test column and road condition column.

4.3 Driving Operation Data Preprocessing

The first 30 rows and last 30 rows of the collected vehicle telemetric data need to be removed. This is because drivers just come out from the parking at the beginning or drive into a parking slot at the end of the driving session. The removed data is always incomplete. Figure 4.6 shows the setting of CSV Reader module of the KNIME Analytic Platform to perform data preprocessing.

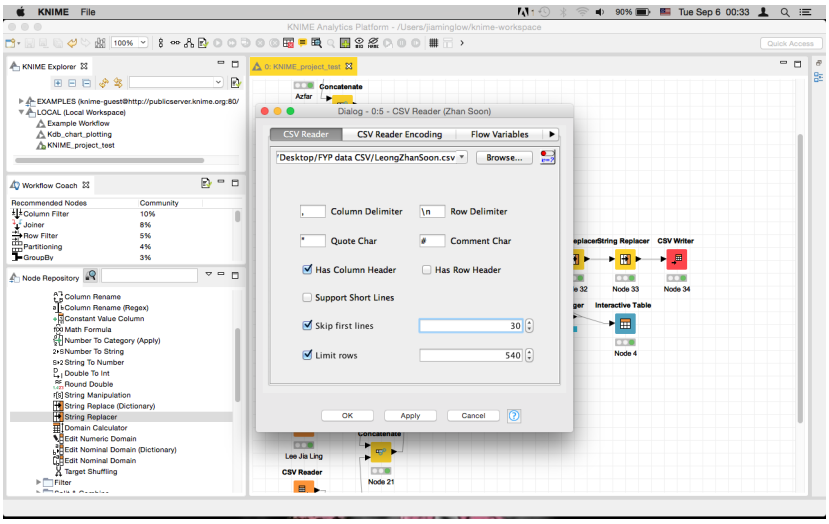


Figure 4.6: CSV Reader settings for data preprocessing.

4.4 Data Fusion

All the pieces of preprocessed data for each driver are required to be concatenated as a big table before performing clustering. In Figure 4.7, Concatenate module provided by the platform combines the processed data.

4.5 Establish the driving operation model by K-Means Algorithm

A workflow is designed and implemented by using KNIME Analytic Platform. The workflow will be able to input the vehicle telemetric data file and perform K-Means Algorithm on the dataset. The number of clusters can be 3, 4, 5 or more. However, three clusters will be identified through the workflow in this project. The three clusters will represent the good, medium, and bad vehicle condition data. In the end of this process, each vehicle operation record will be labelled with the good, medium, and bad vehicle condition data.

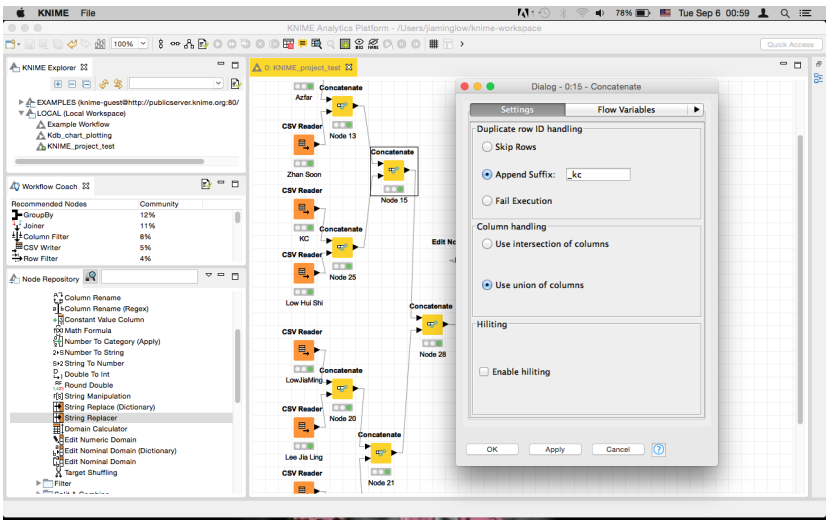


Figure 4.7: CSV files are concatenated by using the concatenation module in KNIME.

4.5.1 Features selection

Some of the features of the vehicle telemetric data are chosen for the usage in clustering. Linear Correlation module is implemented to help features selection. The linear correlations between the features are shown in Figure 4.8. The low saturation of the colors means that the correlation between the two features is low. According to the correlation result, the selected features include GPS Altitude, GPS Bearing, GPS Vehicle Speed (km/h), Engine Speed, Throttle Position, and Speed Test. Engine coolant temperature and acceleration sensor are removed because these two features do not have strong relationship among the rest of the features. The selected features will be used in clustering.

4.5.1 (a) Linear Correlation

Linear correlation is the straight-line relationship between two variables. The range of the value of the linear correlation is between '-1' and '1'. If the value is '1', it means that the relationship is perfect positive relationship. The coordinates of the variables displayed on graph will be similar to a straight line with positive gradient. If the value is '-1', it means that the relationship is perfect negative relationship. The coordinates of the variable displayed on graph will be similar to a straight line with negative gradient. The '0' means no straight-line relationship between the two

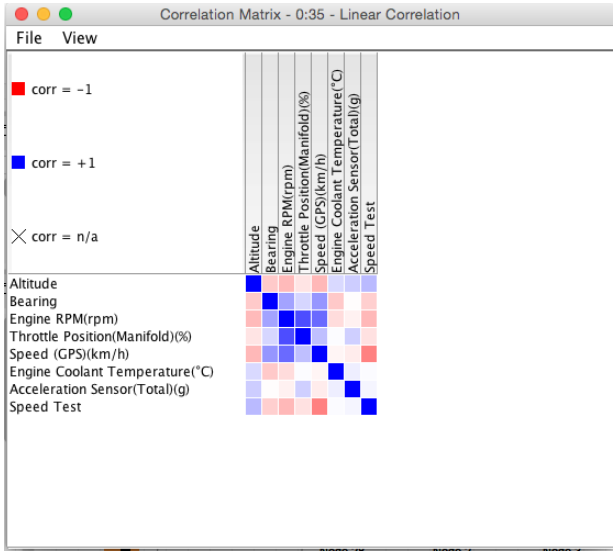


Figure 4.8: The correlation among the features

variables.

4.5.2 Cluster Labelling

After the K-Means Algorithm module execution, three clusters will be generated. Figure 4.9 shows the center coordinate of each cluster. An assumption is made at this point for giving the value to the clusters. According to the coordinates, speed test value is considered to label the clusters. 'cluster_0' is assumed as good vehicle condition data as the speed test value of the coordinate is near to '1'. Most of the driving sessions are observed. Most of the drivers drove the car as normal and did not have dangerous actions performed. So, 'cluster_2' is assumed as medium vehicle condition data due to the larger amount of the records under this cluster. 'cluster_1' is assumed as bad vehicle condition data.

String Replacer module provided by KNIME Analytic Platform is implemented to change the 'cluster_0', 'cluster_1' and 'cluster_2' to 'good', 'medium' and 'bad' accordingly.

Tableau is implemented to show the relationship between the vehicle speed, speed test and the three clusters in Figure 4.10. Based on the graphical result, if speed

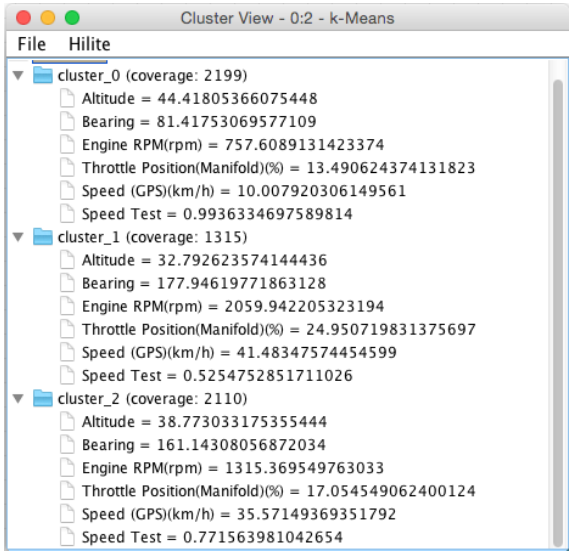


Figure 4.9: The coordinates of cluster center for each cluster.

test value is '-1', the particular record will not be guaranteed to label as medium or bad. The low vehicle speed does not mean that the driving behaviour is good.

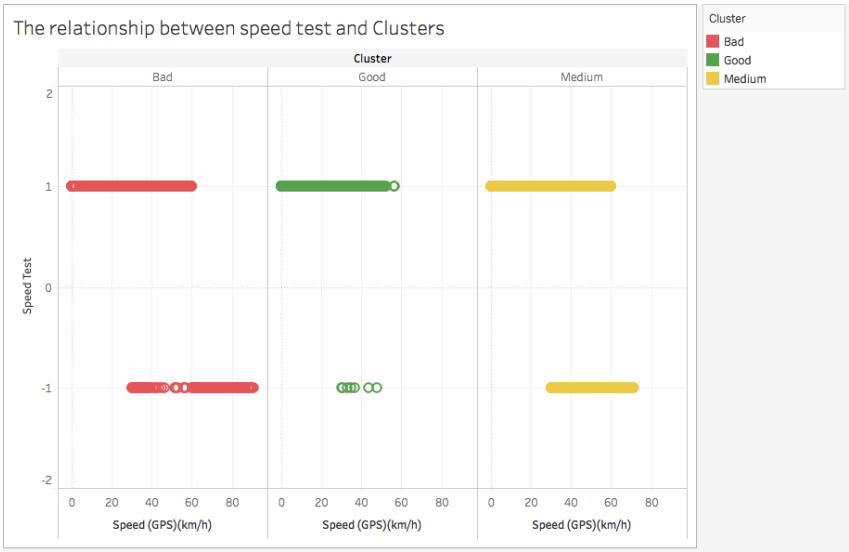


Figure 4.10: The relationship between the speed test and the three clusters.

4.5.3 Workflow Design

The completed workflow is displayed in the Figure 4.11. The CSV Reader is the module that allows CSV file to be inserted into the workflow and perform the data preprocessing at the same time. The Concatenate module will combine the drivers' vehicle telemetric data. The Linear Correlation module is implemented to select the

correct features. The K-Means Algorithm module will perform clustering to categorize the data into three groups. The number of groups is set in the module configuration. Figure 4.12 shows the configuration of the K-Means Algorithm module.

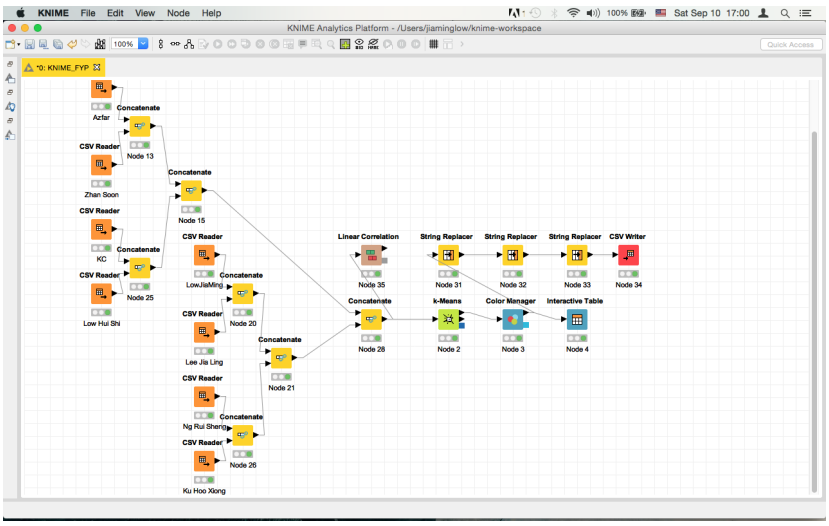


Figure 4.11: The completed workflow for classifying the driving operation model.

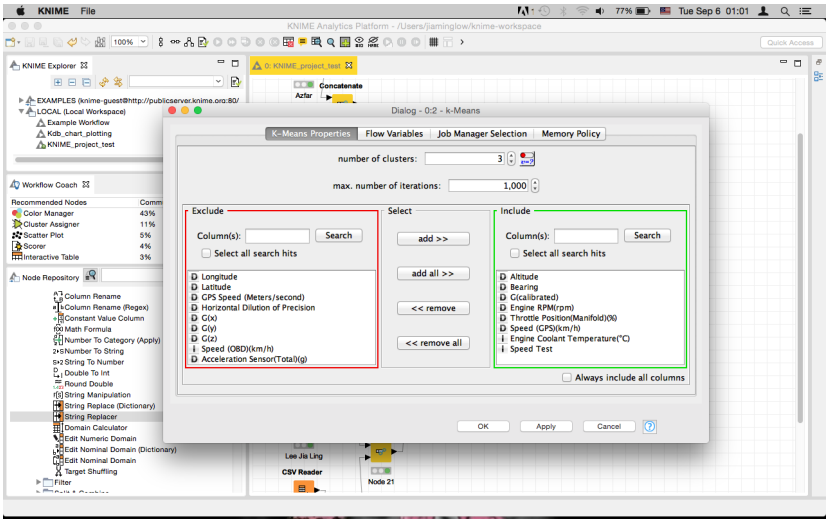


Figure 4.12: K-Means Algorithm module setting.

4.6 Driver Driving Behaviour Profiling

In this project, the driver driving behaviour profile can be determined by some condition. If the total of good VCD is greater than the total of medium VCD and bad VCD at least 10% of the total records, the particular driver will be categorized as low risk driver. If the amount of good VCD do not fulfil the condition, the total of medium VCD and bad VCD will be chosen for classifying the driver. The total of medium

vehicle VCD should over at least 55% of the total medium VCD and bad VCD, and then the driver will be categorized as medium risk driver. Otherwise, the driver will be categorized as high risk driver.

After trying few methods to determine the driver driving behaviour profile, this method is the most suitable among the method. It is because there are two comparisons to be performed to get the driver driving behaviour profile. The first comparison will determine whether the driver is low risk driver. In order to fulfil the condition of the first comparison, the good VCD should over 55% of the total records. If not, the rest of the records will be used for determining whether the driver is medium risk or high risk driver. The medium VCD is required to have 55% among the medium VCD and bad VCD, then the driver only can be categorized as medium risk driver.

The table 4.6 shows the result of the amount of driving records for each cluster and the driver driving behaviour profile.

Driver	Good Model	Medium Model	Bad Model	Driver Driving Profile
Candidate 1	239	138	249	High Risk
Candidate 2	178	120	271	High Risk
Candidate 3	178	156	237	High Risk
Candidate 4	372	207	67	Low Risk
Candidate 5	553	294	95	Low Risk
Candidate 6	144	446	88	Medium Risk
Candidate 7	372	304	205	Medium Risk
Candidate 8	163	445	103	Medium Risk

Table 4.1: The result of the driver driving behaviour profiling.

One of the ways to determine the driver driving behaviour profile has been tried. The method is using the average of the records among the three clusters to compare with the total amount of each cluster. If the good VCD exceeds the average point, the particular driver will be classified as low risk driver. It is same to every cluster.

This method is not suitable for determining the driver driving behaviour profile because there will have same cases difficult to classify the drivers' risk level. In Figure

4.13, Candidate 1 will be hard to classify the driver's risk level as the good VCD and bad VCD exceed the average at the same time. So, this method is not suitable.

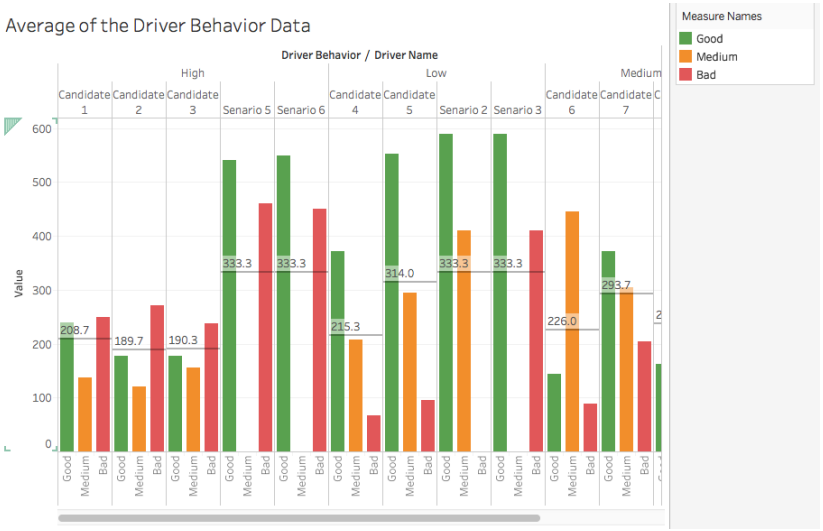


Figure 4.13: The result of the driver driving behaviour profiling by using Tableau side-by-side bar with average point.

CHAPTER 5

IMPLEMENTATION PLAN

5.1 Project Problem Encounter

In this current phase of the project, the route of drivers drove through is inconsistent. It caused the speed limit condition is hard to be determined. It needs a lot of time to check the road condition and the speed limit at the particular path of the route by using Google Street View and Google Earth. The usage of speed limit Application Programming Interface (API) is not free. It is also unavailable to use without an Asset Tracking License.

5.2 Data collection

In the next phase of the project, three routes will be selected. The drivers will be requested to drive through all the routes. After the driving session, a proper questionnaire will be required to fill up by the driver. The questionnaire will ask about the driver basic detail and some driving history or background.

At least a month will be spent to collect the data in the next project phase. In order to train the machine for classifying the driver driving behaviour, the amount of the data collected in the current project phase is still not enough.

5.3 Data Analysis

Since the paths for driving session is consistent, the work for labelling the road condition will be reduced. A proper speed limit labelling will able to be implemented. For the speed test feature, some cases are allowed for drivers to exceed the speed limit. For example, Driver A wants to overtake Driver B's car as Driver B is driving dangerously or the Driver B is drunk. The Driver A behaviour will be considered as

low risk, although Driver A exceeded the speed limit. A lot of situations need to be considered and identified in the next phase of project.

5.4 Machine Learning Technique Implementation

Naive Bayes algorithms will be implemented in the next phase of project. The machine will be trained to classify the driver driving behaviour by using the Naive Bayes algorithms.

5.4.1 Naive Bayes classifier

Naive Bayes is one of the most efficient and effective machine learning algorithms. It is a supervised learning algorithm. It assumes that the features are independent. Each feature contributes to the probability to identify the classification independently. For example, a fruit can be classified as an apple based on the red color, circle shape, and the 6*cm* in diameter. Although the features exist relationship among themselves, the features contribute independently in the probability that to identify the fruit is apple(Ray, 2015).

5.5 Project Plan of next project phase

In the next project phase, data collection is required a lot of time to be done. The papers related to Naive Bayes algorithms are required to review in order to improve the understanding of the algorithms. The next project phase timeline is shown in Table 5.5.

Task \Week	2	3	4	5	6	7	8	9	10	11	12
Literature Review	■	■	■								
Data Acquisition	■	■	■	■	■						
Data Analysis using KNIME					■	■					
Driver Driving Behaviour Profiling							■	■			
Data Training using KNIME								■	■		
Analysing Experimental Result									■	■	
Documentation and Report									■	■	■
Report Submission											■

Table 5.1: Project timeline of Final Year Project II for Trimester 2 2016/2017

CHAPTER 6

CONCLUSION

6.1 Introduction

In the current project phase, normal driver driving behaviour profiling is achieved by using the KNIME. However, the method described in this project needs to be enhanced. A model should be trained for machine learning. In the future, a new driver vehicle telemetric data will be able to be classified without clustering.

6.2 Conclusion

Driver driving behaviour profiling using vehicle on board diagnostic (OBD) information and K-Means Algorithm is described in this project. OBD interface and smartphone are utilized to collect vehicle telemetric data and GPS data. According to the value of linear correlation among the features of the vehicle telemetric data, some of the features are selected for the usage in K-Means Algorithm clustering. The selected features contain GPS Altitude, GPS Bearing, GPS Vehicle Speed (km/h), Engine Speed, Throttle Position, and Speed Test.

Each vehicle operation record is labelled as good, medium or bad condition according to the clusters. Each driver will be categorized to three groups based on the number of each condition the driver had. In the next project phase, this method will be implemented in order to train the machine to classify the drivers.

APPENDIX A

MEETING LOG

REFERENCES

[1] Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., ... Wiswedel, B. (2007). KNIME: The Konstanz Information Miner. In *Studies in classification, data analysis, and knowledge organization (gfkl 2007).* Springer.

[2] Chen, S.-H., Pan, J.-S., & Lu, K. (2015, March). Driving behavior analysis based on vehicle obd information and adaboost algorithms. In *Proceedings of the IMECS international multiconference of engineers and computer scientist 2015* (Vol. 1, p. 102-106).

[3] Kamaruddin, N., & Wahab, A. (2010, June). Driver behavior analysis through speech emotion understanding. In *Intelligent vehicles symposium (iv), 2010 ieee* (p. 238-243). doi: 10.1109/IVS.2010.5548124

[4] Kulanthayan, S., Phang, W., & Hayaticoncatenate, K. (2007, March). Traffic light violation among motorists in malaysia. In *Iatss research* (Vol. 31, p. 67-73).

[5] Miyaji, M., Danno, M., & Oguri, K. (2008, June). Analysis of driver behavior based on traffic incidents for driver monitor systems. In *Intelligent vehicles symposium, 2008 ieee* (p. 930-935). doi: 10.1109/IVS.2008.4621130

[6] Ng, A. (2016). *Unsupervised learning.* Retrieved from `https://www.coursera.org/learn/machine-learning`

[7] Ray, S. (2015). *6 easy steps to learn naive bayes algorithm (with code in python).* Retrieved from `https://www.analyticsvidhya.com/blog/2015/09/naive-bayes-explained/`

[8] Sathyanarayana, A., Boyraz, P., & Hansen, J. H. L. (2008, Sept). Driver behavior analysis and route recognition by hidden markov models. In *Vehicular electronics and safety, 2008. icves 2008. ieee international conference on* (p. 276-281). doi: 10.1109/ICVES.2008.4640874

[9] Stupp, C. (2015). *Consumer groups expose manufacturers for collecting data from connected cars.* Retrieved from `https://www.euractiv.com/section/digital/news/consumer-groups-expose-manufacturers-for-collecting-data-from-connected-cars/`