Design and Development of a Smart Used Car Recommendation System

By

Te Kay Hoe

A REPORT

SUBMITTED TO

Universiti Tunku Abdul Rahman

in partial fulfillment of the requirements

for the degree of

BACHELOR OF INFORMATION SYSTEMS (HONOURS)

BUSINESS INFORMATION SYSTEMS

Faculty of Information and Communication Technology

(Kampar Campus)

JUNE 2025

COPYRIGHT STATEMENT

ACKNOWLEDGEMENTS

I express my deep gratitude to Dr Abdulkarim Kanaan Jebna, my committed supervisor, for his important guidance and assistance during the development of the " Design and Development of a Smart Used Car Recommendation System" project. The project's direction and success have been greatly influenced by Dr Abdulkarim's experience and mentorship, which has played a crucial role in advancing my pursuit of excellence in web application development.

I express my heartfelt gratitude to my family for their unwavering faith in my capabilities and their unflinching assistance throughout this project. Their consistent support has served as a constant source of motivation, empowering me to overcome obstacles and achieve the objectives.

I want to express my appreciation to all persons who have contributed to the success of this project, whether by their assistance, counsel, or motivation. The combined efforts of all of you have played a crucial role in the successful completion of this project, and I am sincerely grateful for your valuable contributions.

Thank you all for your invaluable support and guidance.

ABSTRACT

This project focuses on developing a recommendation system dashboard project for academic purpose, specifically within the field of machine learning. The primary goal is to implement recommendation system algorithms that provide personalized used car recommendations for Malaysian users. This project is required as Malaysia faces a few challenges in the used car market. These problems include choice overload overwhelming users and lack of used car recommendation system tailored to Malaysians looking for used cars. To understand and address these challenges, reviews of past studies related to Malaysian used car market and recommendation system has been done. As a result, this dashboard will use combination of few approaches such as content filtering and clustering to ensure accurate recommendations to users. Technologies that are commonly used in recommendation systems such as Python will also be used for the development process. This project aims to contribute to local used car market by ensuring transparency and healthy interaction between buyers and authentic used car dealers. The output of this project should be able to collect preferences from user, then provide recommendations based on these preferences.

Area of Study: Machine Learning

Keyword: Data Mining, Data Clustering, Used Car Recommendation System, Dashboard, Python

TABLE OF CONTENTS

Contents

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

CRISP-DM	*Cross Industry Standard Process for Data Mining*
BERT	*Bidirectional Encoder Representations from Transformers*
LSTM	*Long Short-Term Memory*
PROMETHEE	*Preference Ranking Organization Method for Enrichment Evaluations*
DB	*Davies-Bouldin Index*
CH	*Calinski–Harabasz Index*

CHAPTER 1

Introduction

The heavy digitization in today's world has transformed the way consumers search and purchase products. One industry that has been the impact of digitization is the used car industry. Traditionally, consumers relied heavily on dealership recommendations or recommendations by people around them. Today, with the increased accessibility to the Internet, the method has evolved into browsing car listings on internet. The advent of Internet indeed provides more advantage than traditional methods such as saving time and energy. However, Internet can also be a double-edged sword.

With thousands and thousands of used car listings being uploaded to the internet daily, consumers are constantly bombarded with tremendous amount of car information to consider about. Although there are dedicated platforms for used car market such as Carlist, Carsome and Mudah, it only offers consumers access to used car listings, which still leaves users with questions such as "Which is the car that best meets my requirement and budget?". As a result, consumers often face a dilemma in choosing a used car on the internet. This presents us with a problem named choice overload, a scenario where there are too many available options, hindering a consumer's ability to make a choice. Consumers that are trapped in this problem may be overwhelmed, leading to indecision or uninformed decision.

In Malaysia, this problem is further amplified due to the booming used car market. Statistics show that in 2024, the estimated project revenue in passenger cars market in Malaysia is US$16 billion [1]. By 2028, the revenue is expected to experience a compound annual growth rate of 3.05% [1]. As of 31st July 2024, the total number of registered cars in year 2024 has reached approximately 480,000 [2]. Malaysia even became the second-largest auto market in Southeast Asia, overtaking Thailand [3]. All these signs tell us that the car market in Malaysia is growing, until a point where the registered vehicle population (36.3 million) has exceeded human population (32.4 million), as in 2023 [4]. Among the registered vehicle population, almost half of it is comprised of cars at 17,244,978. Meanwhile, the used car market is also growing and is predicted to reach MYR 26.2 million by 2027 [5]. The causing factors are believed to be sales tax exemptions, expanding middle class, and the rise of trade-in culture [5].

Our used car recommendation system is designed to address this problem, providing personalized recommendations tailored to users' personal preferences and lifestyle needs. The system leverages modern algorithms and user-centric approaches to provide a platform prioritizing user experience. Our user-friendly interface design guides users step-by-step in their decision-making, enabling them to make decisions in a comfortable, hassle-free environment.

By filtering through extensive data obtained from dealership listings on Carlist, our system simplifies the decision-making process when purchasing a used car. Instead of being presented with overwhelming amounts of available car options, users will be provided with a concise list of suitable used car options that are tailored to their preferences through a series of selection.

Ultimately, we want our used car recommendations system to empower consumers with confident in making decisions when purchasing a used car market, leading to a satisfactory experience. By providing a smooth decision-making process and carefully selected recommendations, we hope that users can explore the available options effortlessly.

1.1 Problem Statement and Motivation

The used car market size in Malaysia has been growing over the past decade, driven by economic, technology and policy. As of 2025, the market size is projected to achieve USD 26.5 billion, with a CAGR (Compound Annual Growth Rate) of 5.4%, that is expected to reach USD 34.47 billion in 2030. The expected growth of USD 8 billion over the five-year period has certainly indicated the Malaysian's used car market's potential and increasing importance in the Malaysia automotive industry.

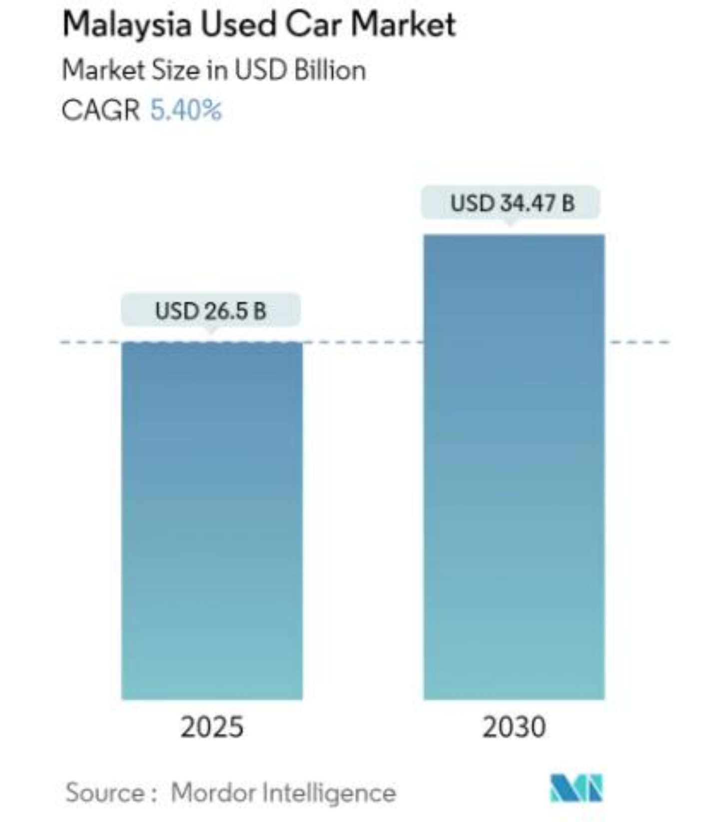


Figure 1. Expected Used Car Market Size in Malaysia [6]

This sign of rapid expansion indicates a rising issue which is: increasing choice complexity for consumers. This is because with more vehicles are entering and available on platform for buying and selling, the amount of available used car listings is becoming too much, that it is overwhelming for consumers. Not to mention that online sales account for the largest market share in Malaysian automotive industry, automotive consumers are constantly being bombarded with tremendous amounts of information. This requires consumers to take a lot of information into account, ranging from used car specifications to the seller's reputation, which leads to choice overload [8]. This choice overload can not only affect the consumer's decision quality but also increase the chances of consumers feeling dissatisfied with the purchase [9].

There is also a lack of used car recommendations system tailored to Malaysian consumers. Majority of the car recommendation systems such as AutoTrader or Carvana might be available, however they are tailored towards the Western consumers. Since domestic car brands such as Proton and Perodua are dominating the Malaysian automotive industry [7], AutoTrader and Carvana may exclude Malaysian domestic car brand providing recommendations that are irrelevant for Malaysian consumers, given that preference for domestic car brands has been deeply engraved in Malaysian's car purchasing culture. There is a famous Malaysian car recommendation system called CarBase.my, however the system primarily focuses on brand new car, highlighting a need of used car recommendation system that is catered to Malaysian consumers.

Consequently, there is a need for a used car recommendation system catered towards Malaysian consumer that filter out irrelevant listings, presenting consumers with a concise list of used car recommendations that matches their preferences after series of carefully selection. This approach can enhance consumers' decision quality and satisfaction, as it has been proven in Lepper and Iyengar's study, where participants reported better satisfaction when they have fewer options to choose from [9].

1.2 Project Objective

To develop a model that leverages machine learning technology to provide used car recommendations for Malaysians who are looking for used cars to buy. The aim is to enhance Malaysian used car consumers' experience when buying them. The system will aid the Malaysian used car consumers by providing them with relevant information based on preferences of users, which they have specified and filtered using the system. This can help providing them with useful insights when doing consideration, hence making informed decisions.

1. To identify a suitable machine learning algorithm in providing used car recommendations

2. To develop and train a model that can give relevant recommendations to user

3. To design and implement a dashboard interface that allows users to enter their inputs of desired used car specifications

1.3 Project Scope

The scope of this project is to develop a dashboard that implements machine learning model to provide used car recommendations to Malaysians who are looking for used cars. Our dashboard can empower their ability to look for used car tailored to their preferences. The dashboard will be using cleaned data, which we have obtained from web scraping, particularly from Carlist.my, one of the prominent automotive websites in Malaysia. The dashboard will take factors provided by users such as budget, car brand preferences and mileage into account when providing recommendations, ensuring that the results tailored and highly relevant to user needs. This project is focused on Malaysia region, aiming to provide a localized used car recommendation system, benefiting local players in the used car market.

1.4 Contribution

The contribution of this project is significant, especially towards Malaysian consumers and automotive industry. It first addresses the distinct Malaysian used car ecosystem by introducing a localized used car recommendation system, catered towards the needs, preferences and behaviours of Malaysian consumers, which other car recommendations system like AutoTrader could not do. For instance, this system includes Malaysian domestic car brands such as Proton and Perodua, which are often consumers choice when going for a used car. This project also ensures that the information given to users are updated. To ensure the relevancy of information provided, the dataset used for recommendations will also undergo regular updates.

1.4 Report Organization

This report consists of 7 chapters, each outlining different sections of the project. In Chapter 1, a comprehensive background of the project, outlining the problem statement, project motivation, scope, objective, contribution and report organization. In Chapter 2, a literature review comprises of studies that have been done in the past to explore recommendations system will be provided, highlighting their approaches, strengths and weaknesses. Chapter 3 discusses the methodology to be used for this project, explaining each phase of the methodology and acting as a guideline in developing the project. Chapter 4 presents the system structure and design. Chapter 5 explains the system implementation process, outlining the effort done in developing the project. Chapter 6 evaluates the project, determines if the project achieves the objectives. Finally in Chapter 7, a conclusion of the project will be given, along with some recommendations for future improvements.

CHAPTER 2

Literature Review

A comprehensive review related to past research and existing platforms in the field of car recommendation system will be provided in this chapter. The main objective of this chapter is to explore and examine different methods used in previous studies, to gain a deeper understanding of characteristics of each method, and to analyse their strengths and weaknesses. From this review, a clearer direction for developing a used car recommendation system tailored to Malaysian user can be provided.

2.1 Algorithms Used in Car Recommendation Systems

Several techniques have been implemented in the field of recommendation system for cars, ranging from traditional models like knowledge-based models, to modern artificial intelligence-driven models.

Knowledge-based model is one of the methods implemented to build a recommendation system. Alabduljabbar [10] developed a recommendation system using knowledge-based model that provides recommendations based on user-defined criteria such as brand, price and manufacturing year. This method performed well, especially under the situation where there were none prior user interaction data. The model addressed the cold-start problem, where recommendation system cannot provide useful information due to limited information obtained from user, such as browsing history.

Deep learning model on the other hand was implemented by Karthikeyan [11] to build a model that processes unstructured text data and sequential data. The team proposed an Artificial Intelligence framework that processes data such as showroom visits, search history and social media using BERT (Bidirectional Encoder Representations from Transformers) and LSTM (Long Short-Term Memory). BERT will extract semantic meaning from users' text input while LSTM will analyse the sequential data. The strength of this method is that the system understands natural language inputs from users, enabling better personalization using BERT model. However, high computational power, significant training time and high training cost is a big downside of this method, making it cost-ineffective for small developers.

Another approach studied is the PROMETHEE method (Preference Ranking Organization Method for Enrichment Evaluation) [12]. PROMETHEE is a decision-making technique that determine the order of priority in a multi-criteria decision-making process. The method has proven successful track record of providing accurate recommendations for air conditioners, student scholarships and others. The method is also able to handle both quantitative and qualitative data effectively. In their car recommendation system, the author employed PROMETHEE to rank car attributes based on quantitative and qualitative criteria obtained from user survey and web scraped data. The satisfaction rate of this method reached 89.2%, however has a weakness. The model cannot adapt to changing user requirements and require manual modifications to provide accurate recommendations again.

Another simpler method introduced is the Decision Tree Classifier. Pawar [13] used car attributes such as price, transmission and mileage to develop a model that has reach an accuracy rate of 86.43%. The simplicity of Decision Tree allows easier understanding of visual structures and uses low computational resources when compared to other methods. LabelEncoder and hyperparameter tuning via grid search were also implemented to further optimize the model's performance. Specifically, LabelEncoder converts categorical data into numerical data for better understanding. However, the method may be less powerful in predicting, compared to other more advanced models when dealing with more complex user behaviours.

2.2 Review of Existing Online Tools

Aside from studying the articles, literature review has also been done on existing car recommendations system exist online. One notable example is Carwow [14], an online car reviews site based in Europe that features a car recommendations system to assist users in finding a car that matches their preferences. The system uses a filter-based recommendation system where users select desired car attributes from a fixed dropdown menu list. The system then filters and returns results that matches user inputs.

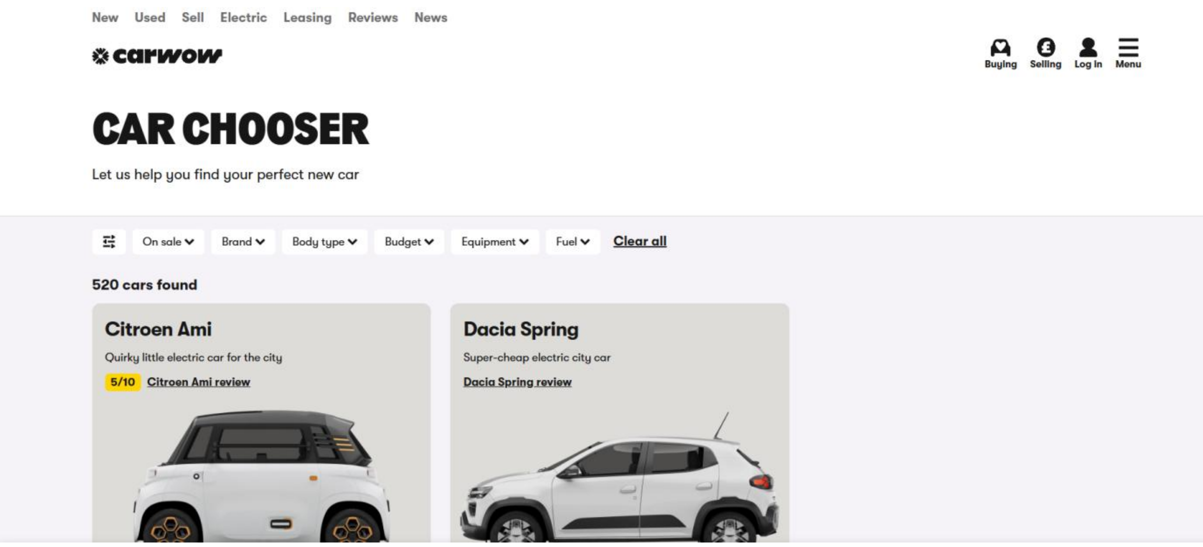


Figure 2. Main Menu of Carwow Car Chooser

One key advantage of Carwow is its clean and intuitive interface. The visually appealing design allows users to navigate through the system with ease. Upon loading the website, users are presented with a full list of available cars. Users can then adjust their filters to refine their selections. The website also has collaborations with dealerships and manufacturers, offering users access to essential car data such as real-time price, available discount and more. Additionally, there are also reviews given by expert car reviewers for each car models. When user clicks on the "Car review" link, they will be redirected to a review page of the selected car. The engagement of the page is also high, which suggests a good reputation of the website.

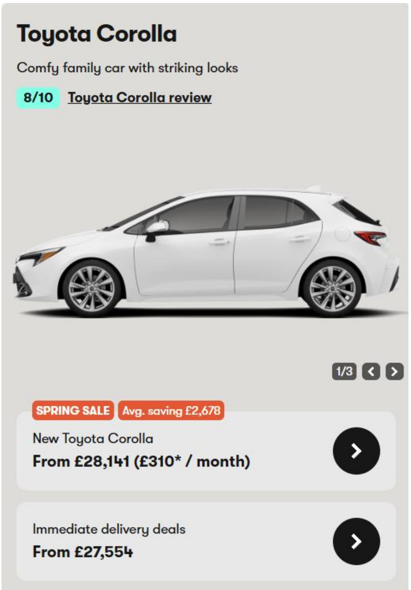


Figure 3. Recommended Car Details

Figure 4. View Counts of the Review Page

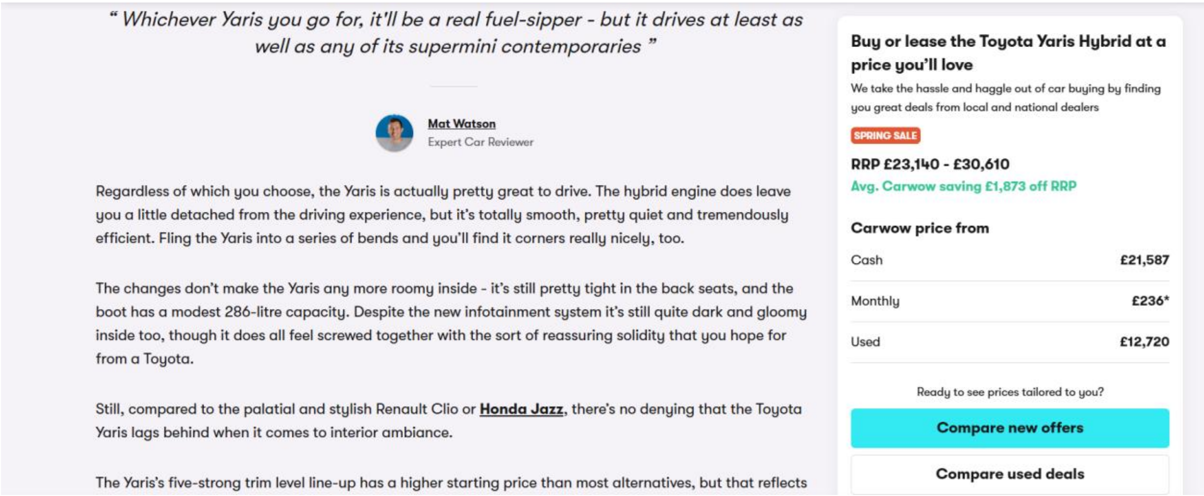


Figure 5. Review of an Expert Car Reviewer

However, Carwow has a potential drawback which is the potential bias in its recommendations due to its partnership with commercials. Since Carwow partners with specific brands, the recommendations provided might favour these brands, resulting in a less personalized recommendation. The lack of predictive analytics also limits its ability to provide refined recommendations based on user behaviour since it only uses manual filtering.

Another existing online car recommender system reviewed is Car Chooser by CarExpert [15], an Australian car website. Similar to Carwow, Car Chooser utilizes manual filtering to recommend cars. Unlike Carwow which presents users with a full list of available cars, Car Chooser interacts with users immediately by asking a series of questions about their desired car to collect user preferences. This approach is more straightforward than Carwow and can streamlien the selection process.

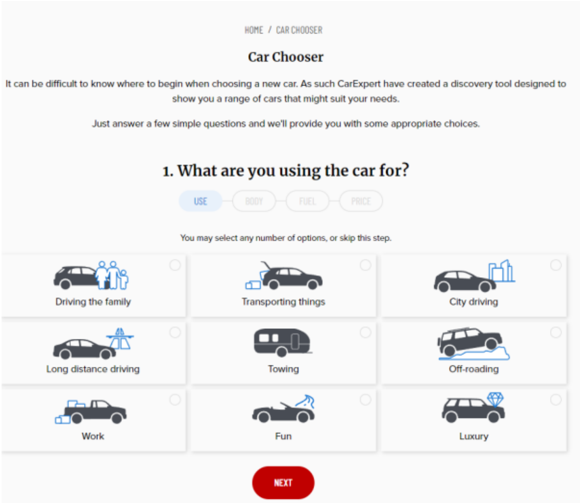


Figure 6. System Asking User Questions to Determine Preferences

After finish collecting the user inputs, the system search matching results and present it to user.

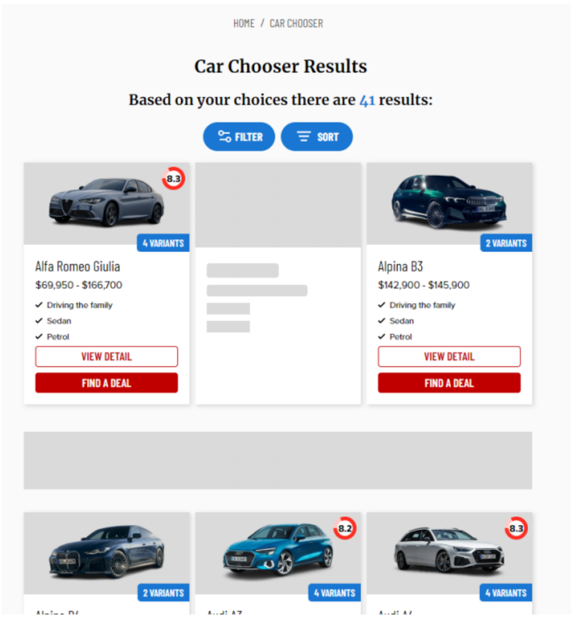


Figure 7. Recommendations Based on User Preference

Similar to Carwow, clicking each car results will redirect users to the cars' respective detail page that shows further information of the car, such as the model variant, fuel efficiency and more.

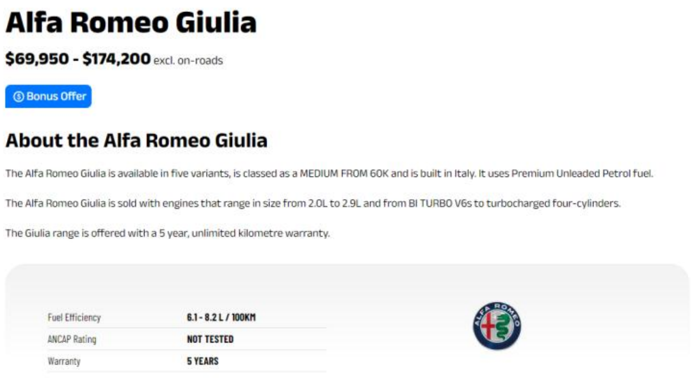


Figure 8. Car Details Page

One advantage of Car Chooser that sets it apart from Carwow is its car comparison tool. This feature allows user to compare two cars side-by-side, comparing the attributes such as price range, technology supported and fuel capacity.

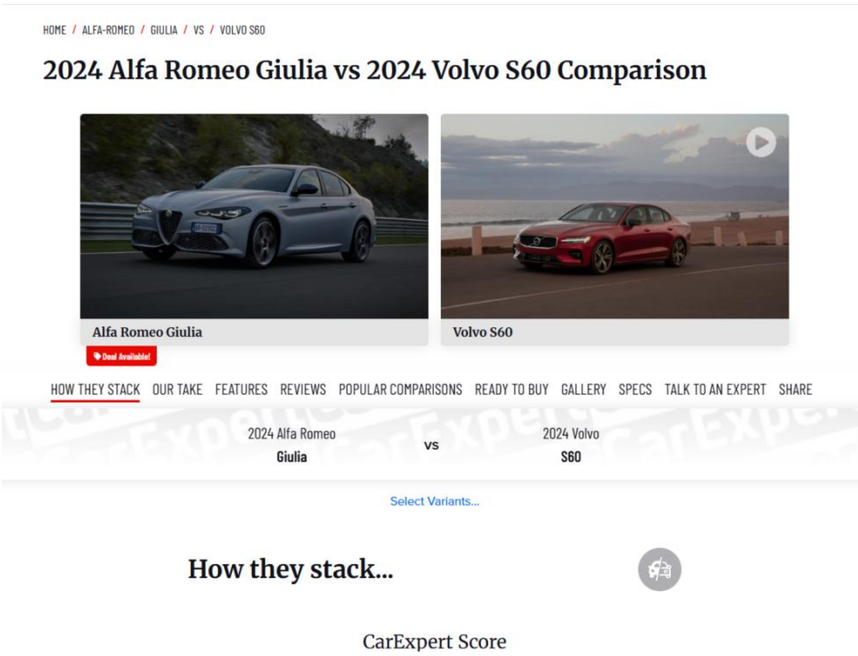


Figure 9. Car Comparison Tool Part 1

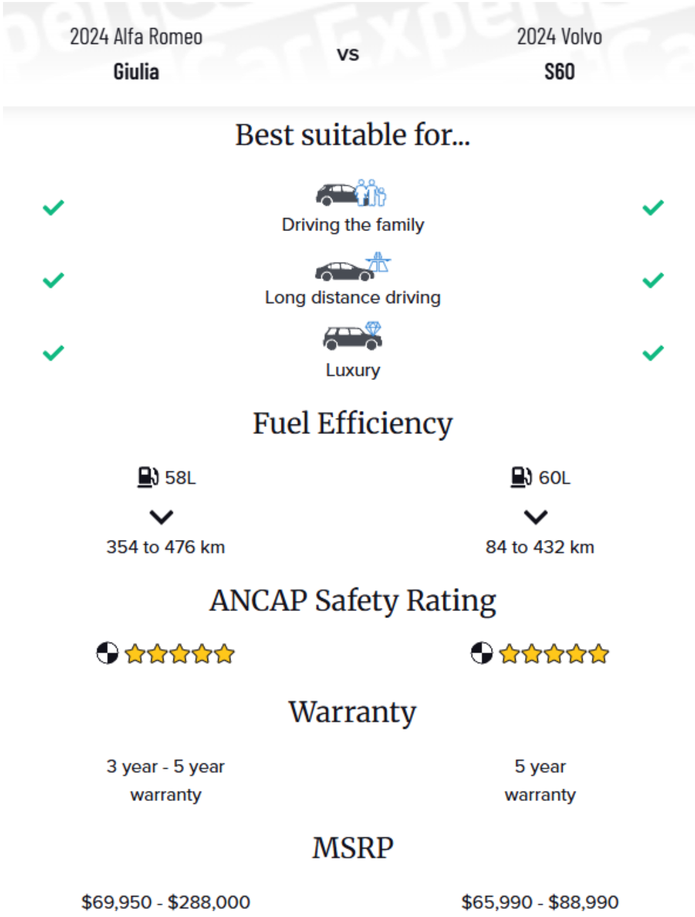
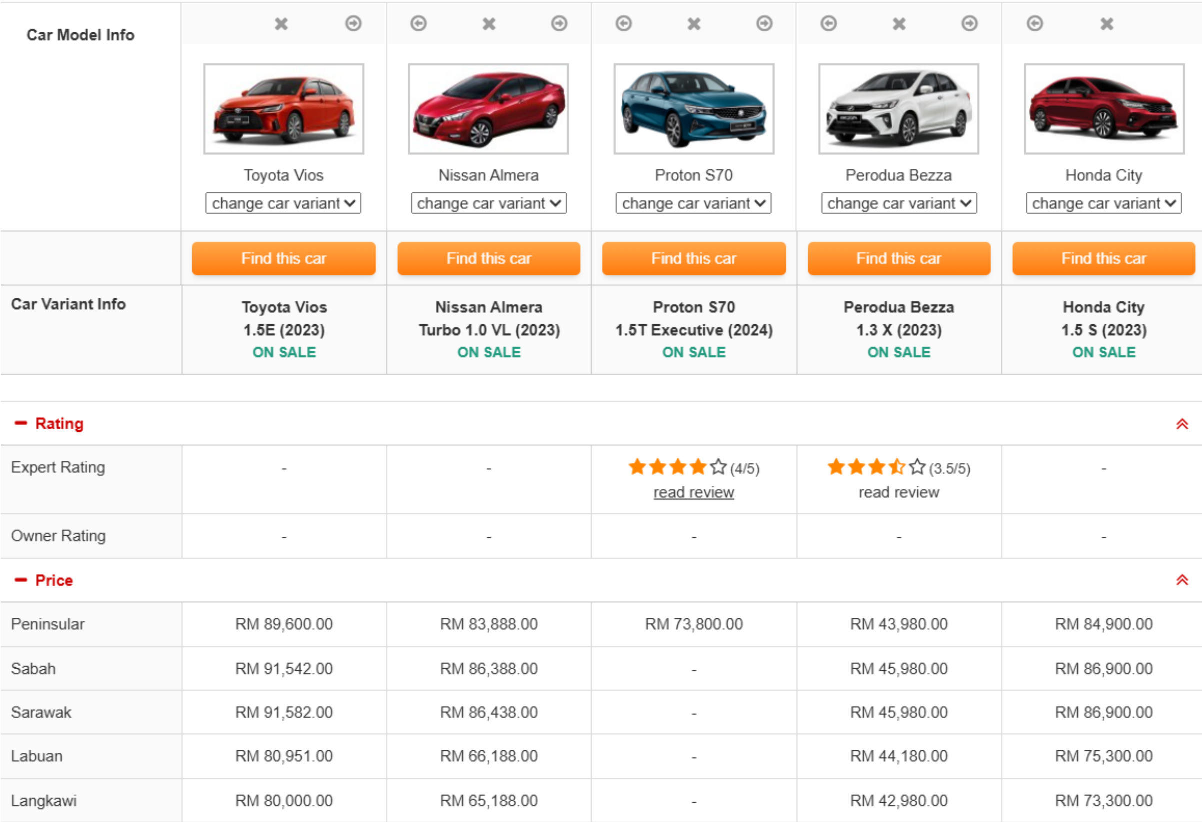


Figure 10. Car Comparison Tool Part 2

While Car Chooser shares many advantages with Carwow, it also faces similar disadvantages, which have been discussed above.

CarBase.my [16] is a car recommendations system presented by paultan.org, a famous automotive website based in Malaysia. To the best of our knowledge, CarBase.my is the only famous car recommendation system that is based in Malaysia. There is also a possibility that other similar system that are based in Malaysia exists but were not found during our study.

CarBase does what other existing platform does, which is to provide car recommendations based on user-defined criteria. It also has car comparison tool that allows users to compare up to 5 cars. There is also community engagement within the system, where there are reviews and comments sections that allow car owners to share and exchange their opinions, which is something that we want to create in our study. Similar with Carwow and CarExpert, CarBase also lacks machine learning algorithms, which could enhance the recommendations and increasing consumer satisfaction.



Car Model Info	Toyota Vios	Nissan Almera	Proton S70	Perodua Bezza	Honda City
	change car variant	change car variant	change car variant	change car variant	change car variant
	Find this car	Find this car	Find this car	Find this car	Find this car
Car Variant Info	Toyota Vios 1.5E (2023) ON SALE	Nissan Almera Turbo 1.0 VL (2023) ON SALE	Proton S70 1.5T Executive (2024) ON SALE	Perodua Bezza 1.3 X (2023) ON SALE	Honda City 1.5 S (2023) ON SALE
Rating					
Expert Rating	-	-	(4/5) read review	(3.5/5) read review	-
Owner Rating	-	-	-	-	-
Price					
Peninsular	RM 89,600.00	RM 83,888.00	RM 73,800.00	RM 43,980.00	RM 84,900.00
Sabah	RM 91,542.00	RM 86,388.00	-	RM 45,980.00	RM 86,900.00
Sarawak	RM 91,582.00	RM 86,438.00	-	RM 45,980.00	RM 86,900.00
Labuan	RM 80,951.00	RM 66,188.00	-	RM 44,180.00	RM 75,300.00
Langkawi	RM 80,000.00	RM 65,188.00	-	RM 42,980.00	RM 73,300.00

Figure 11. Car Comparison Tool on CarBase

What sets it apart from Carwow and CarExpert is that it focuses on Malaysian car. However, the system also primarily focuses on new car, catering their services towards consumers who are interested in getting a brand-new car. Besides that, paultan.org, the organization behind CarBase.my is a reputable automotive website in Malaysia. Therefore, the expert reviews given in CarBase are trustworthy.

Figure 12. paultan.org the Organisation Behind CarBase

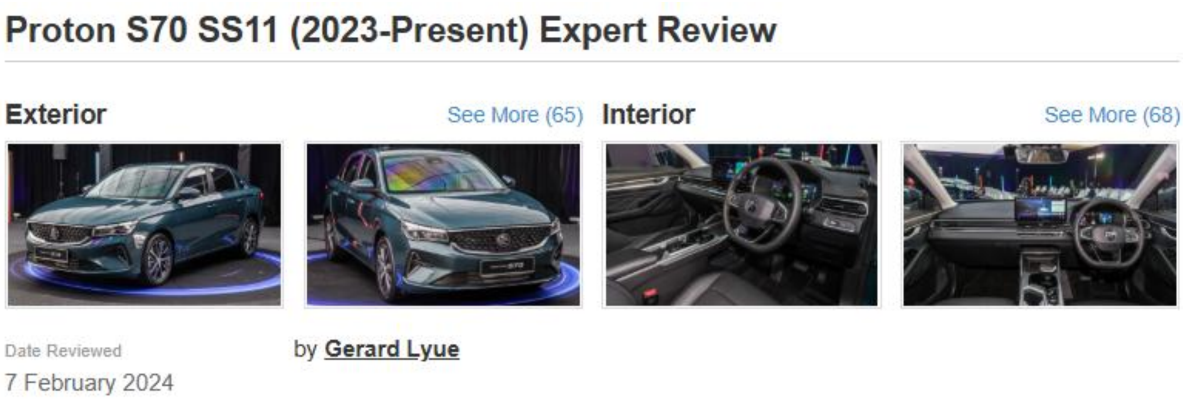


Figure 13. Review from Expert in paultan.org

Summary of Articles/Projects Reviewed:

Table 1. Summary of Literature Review

Author(s)	Objectives	Dataset	Algorithm	Results	Contributions
Alabduljabbar et al. (2023)	To compare item-item, user-user and knowledge-based models	“Driver Vehicle Module (DVM) Car “ on Figshare	Item-item, user-user and knowledge-based filtering	Knowledge-based outperformed other models	Proved knowledge-based model’s ability to address cold-start problem with accurate recommendations
Karthikeyan et al. (2025)	To develop a car recommendation system using textual and sequential data	-	BERT and LSTM (Deep Learning)	Process natural language text input effectively	Innovative approach to process unstructured data (text) for recommendations
Mahgfuri et al. (2022)	To apply PROMETHEE method for recommendations requiring multiple criteria	Web scraping	PROMETHEE	89.2% user satisfaction	Displayed PROMETHEE’s strength in recommending items using

Author(s)	Objectives	Dataset	Algorithm	Results	Contributions
					multiple user-defined priorities
Pawar et al. (2024)	To develop a car recommender using Decision Tree Classifier	Public database (did not specifically mentioned which)	Decision Tree, LabelEncoder, Grid Search	86.43% of accuracy	Created a simple yet decent performing model
Carwow	To filter and recommend cars	-	Filter-based recommendation system	Clean UI design, partnership with dealers; but lack advanced analytics	Partnership with commercials
CarExpert Car Chooser	To provide a guided car recommendation through a series of questions	-	Filter-based recommendation system, comparison tools	Intuitive flow, car comparison tool, but lack advanced analytics	Easy to understand process and guided filtering
CarBase.my	To provide recommendations that are based in Malaysia	-	Filter-based recommendation system, comparison tool	Active community	Provide platform for Malaysian consumers who are interested in brand new car

In conclusion, this chapter provides a comprehensive review of the methodologies used for car recommendation systems. Existing tools online were also covered in this literature review. The methodology covered were knowledge-based models, deep learning models, Preference Ranking Organization Method for Enrichment Evaluation and Decision Tree Classifier. Tools reviewed were Carwow, Car Chooser by CarExpert, and CarBase.my.

Knowledge-based model and Decision Tree are cost-effective and easy to use, while Deep Learning and Preference Ranking Organization Method for Enrichment Evaluation can offer further personalized recommendations, albeit complex and advanced. Existing online tools showed similar strengths and weaknesses but also showed us interesting features such as car comparison tool.

To summarize, the literature review has provided a direction for work to be done during the development of the used car recommendations system.

CHAPTER 3

System Methodology/Approach

This chapter introduces CRISP-DM framework, the methodology planned to be employed for the development of this project. CRISP-DM stands for Cross-Industry Standard Process for Data Mining, a framework designed for data mining and analytics projects. The framework provides a structured, standardized approach that breaks the CRISP-DM process intro six phases, namely **Business Understanding**, **Data Understanding**, **Data Preparation**, **Modeling**, **Evaluation**, and **Deployment** [17]. CRISP-DM is one of the most widely used methodology in developing projects under the field of data mining and analytics. The strength of CRISP-DM provides developers with the ability to revisit previous stages anytime they want to make changes, providing flexibility.

3.1 CRISP-DM

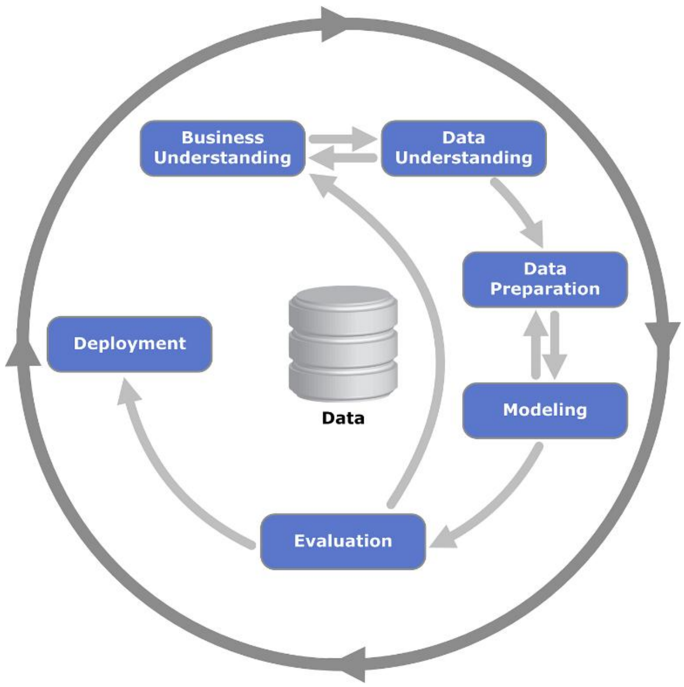


Figure 14. CRISP-DM Phases Diagram [18]

3.1.1 Business Understanding

The first phase focuses on understanding the requirements from users' perspective, then establishing a clear objective, which in our project, is to develop a smart used car recommendations system that provides recommendations based on user preferences such as price, mileage, brand, and more.

To support this objective, the background of Malaysian used car market and ecosystem was studied. This study was crucial to identify few key factors that influence buyer decision, helping the feature selecting process in the next phase and ensure that the system will align with user behaviour in real world scenario.

3.1.2 Data Understanding

In data understanding phase, the data attributes required to achieve the project objectives will be identified and collected. During this phase, the required dataset will be collected using web scraping tool directly from Carlist.my, one of the prominent car listing websites in Malaysia. Selenium will be used during this phase as it mimics human behaviour when scraping data. This is because traditional methods such as Beautiful Soup does not seem to be viable, as Carlist.my seems to be implementing anti-bot mechanisms.

Tons of car attributes, ranging from general specifications like mileage, to more detailed specifications such as brake systems will be collected during the scraping process. Some of the scraped car attributes are, Price, Name, Manufacture Year, Mileage, Transmission, Number of Gears, Engine CC, Engine Type, Aspiration, Fuel Type, Fuel Consumption, Suspension and more. The collected data was stored in an Excel file and is ready to be processed in the next phase.

3.1.3 Data Preparation

This phase is extremely important, as it is critical in transforming raw scraped data into a clean format that is suitable for analysis. A significant amount of time will be spent on this phase to ensure a dataset with high quality. Few tasks such as removing data duplicates, handling missing data values, standardizing units, and more will be carried out to clean and preprocess the data. A few attributes such as Compression Ratio, Suspension, Wheel Size and more are also dropped due to a number of car listings did not cover them. New features derivation may even be used if it proves to be beneficial for the model performance. This goal of this phase is to deliver a well-structured and high-quality dataset for the training of machine learning model. By leveraging a high-quality dataset, the accuracy and performance of the model can be improved.

The dataset will also be splitted into train set and test set for evaluation purpose. The train set will be used during modelling phase, while the test set remains untouched until the evaluation test.

3.1.4 Modelling

During the modelling phase, rule-based filtering system was developed first to create the direct search function. Users will use this feature to key in their preference of the car such as the Color, Brand, etc., while the system takes in the input and provide directly matched results. This direct search feature is widely seen in other tools such as CarWow, CarBase and CarExpert.

The direct search function is also important, as it allows recording of user inputs. The recorded inputs will be used to recommend other similar cars to the user, which involve the use of unsupervised learning techniques, specifically data clustering. The purpose of recommending partially matched results is to pique the users' interest. In the article by Zeno Olech and others [18], it was suggested that recommending "serendipitous" items that share subtle characteristics and are not from the same category can lead to higher user satisfaction. From business perspective, recommending partially matched result can address the trap of the Matthew effect, where less popular item become increasingly less popular due to popular items being recommended too much, resulting in stagnant stock.

To achieve the effect of recommending partially matched results, different clustering methods were tried and compared to select the best method. The methods include K-Means, Hierarchical Clustering and more. These methods group objects within the car dataset based on the attributes, providing clusters that share similar characteristics such as fuel type. When recommending partially matched results, cars from other clusters that are close will be selected and description of why they may be suitable for the users will be shown. After developing all the models, each model will be evaluated using evaluation metrics such as the Silhouette score, Calinski-Harabasz index, and dendrograms to determine the optimal k value for each method. A dashboard is then designed for integration with model in later stage.

3.1.5 Evaluation

At this stage, it is crucial to find out the best clustering algorithm with the best K value, as it will be used to develop the final model. Therefore, the best K value of each model will be compared using evaluation metrics such as the Silhouette score, Calinski-Harabasz. The final best clustering model and K value will be selected to be trained in the next phase.

3.1.6 Deployment

In this final phase, a final model will be trained using the best clustering method and K value. It is then integrated into the dashboard created earlier. The dashboard allows user to interact with the model and obtain recommendations from it.

CHAPTER 4

System Design

4.1 Use Case Diagram

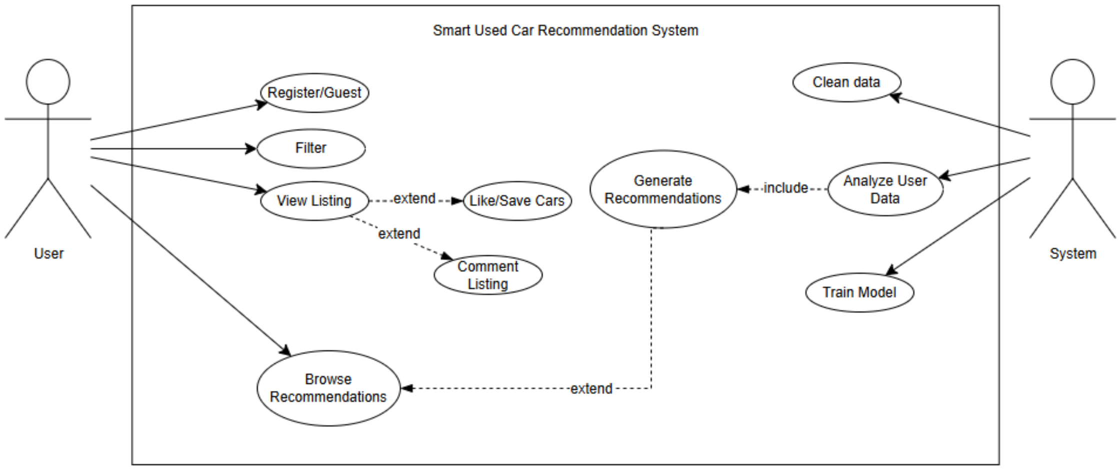


Figure 15. Use Case Diagram

The use case diagram above illustrates how the two actors, “User” and “System” interact. User represents the end users who are using the dashboard, while “System” indicates the functionalities of the dashboard. Firstly, “User” can “Register” and create a profile for the dashboard. If “User” does not want to register a profile, they can choose to continue as a guest, where their data such as browse history will not be collected. Then they can use “Filter” to input their preferences of car to check for car with specific criteria. The “View Listing” allows “User” to check detailed information of the car, where they can also “Comment Listing” or “Like/Save Cars”. These actions will then be saved in their profile, where “System” actor can retrieve and analyse the data. Then the “System” will generate recommendations for “User” to “Browse Recommendations”. “System” actor on the other hand will “Train Model” and conduct “Clean data” task for data preparation.

4.2 System Design

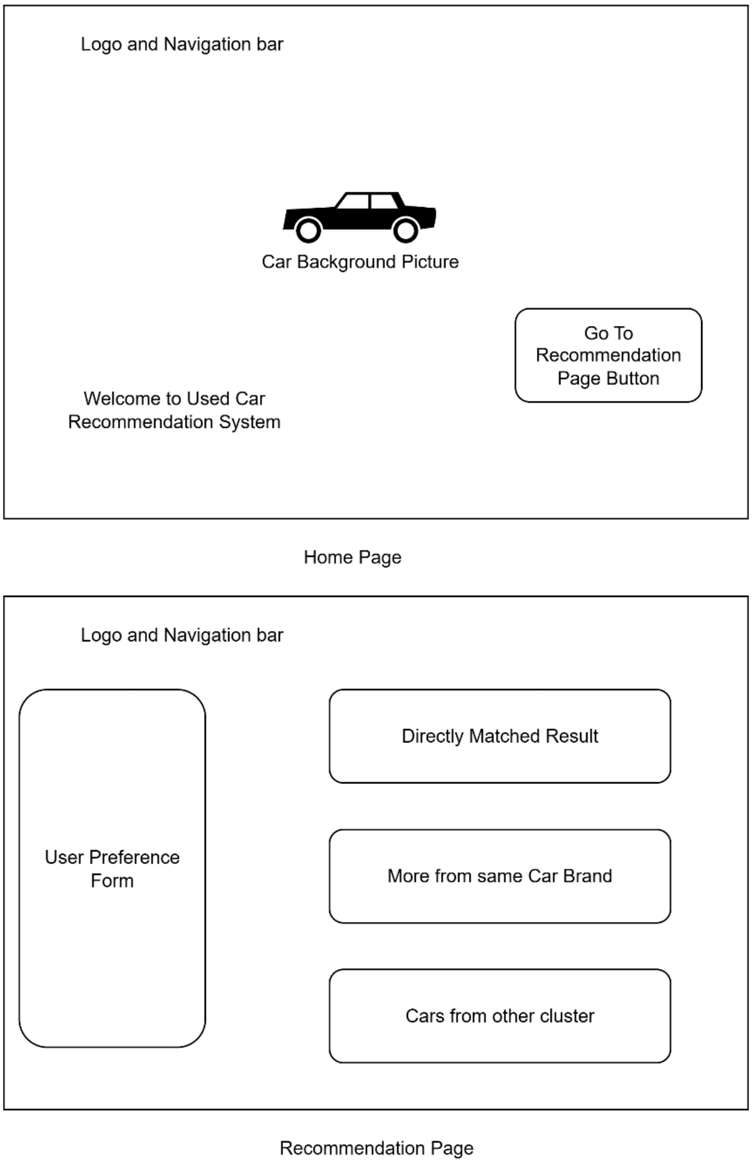


Figure 16 Script Loading Existing Listings

The design for the system consists of two pages, namely Home and Recommendation. Each page has a Logo and Navigation bar that when clicked on, will redirect users to Home Page. When users load up the system, they will be presented with the Home Page. The Home Page sets to give users a good first impression by presenting the system in a visually appealing manner. When the “Go to Recommendation Page Button” is clicked, users will be redirected to Recommendation Page.

At Recommendation Page, users will need to key in their preference inputs at the left panel. After confirming the preference, users will need to click confirm, and will be presented with directly matched results, more cars from same brand and recommendation from other clusters.

CHAPTER 5

System Implementation

Suitable hardware equipment and software requirements are essential for successful development and implementation of a smart used car recommendation system. In this chapter, we will discuss the hardware specifications and software technologies that are selected for development. These items were carefully selected to ensure a functional while appealing system. Reviews and justifications for using these items will be provided. Explanation of how CRISP-DM was implemented will also be provided in this chapter.

5.1 Hardware

The hardware involved in this project is laptop. The laptop will be required for coding, developing, testing, debugging and deploying the dashboard. Data collection, pre-processing and other more are also going to be done using the laptop.

Table 2. Specifications of Laptop

Description	Specifications
Model	FORGE15X 2021 RTX 3060 Edition
Processor	Intel Core i5-10300H
Operating System	Windows 10
Graphic	NVIDIA GeForce RTX 3060 Laptop GPU
Memory	2x 8GB DDR4 RAM
Storage	512GB SSD

5.2 Technology Involved

The following few technologies and software are identified and obtained before the project execution.

Table 3. Technology Used for Project Development

Name	Description
Microsoft Visual Studio Code	Python IDE environment for code compiling and Django project management
Python	To develop the model and dashboard
Selenium	To scrape data and mimicking human browsing behaviour
Django	To deploy app and customize the user interface design

5.3 Setting and Configuration

A successful project requires well integration between components that are carefully selected. In this project, a lot of libraries were leveraged to perform specific activity on the dataset and designing the dashboard interface. In collecting data, processing the data and developing the model, libraries such as “sklearn”, “selenium”, “undetected chromedriver”, “pandas”, “numpy”, “matploblib” and more. “selenium” and “undetected chromedriver” were essential in collecting the data through web scraping as they can automate the page navigation and mimic human browsing behaviour, bypassing the website’s bot detection. The scraped data are then saved using “pandas” feature. The “preprocessing” and “impute” libraries is then used to preprocess and scale the features.

“sklearn” library contains a lot of machine learning libraries that are crucial for model development. Libraries such as “AgglomerativeClustering”, “KMeans” and more were leveraged for model development. “matplotlib” library was also used to generate plot graph and visualize the data for better interpretation.

“django” library was also important in designing the dashboard interface and defining the application’s business logic. The Model-Template-View structure of Django allows separation of different layers, making the development process easier since the the layers are independent.

For instance, the interface design was done in Template, while the business logic was coded in View. “render” library is then used to load the dashboard.

These libraries were then integrated together to create a dynamic used car recommendation system.

5.4 CRISP-DM

5.4.1 Data Understanding

The data used for this system is sourced from Carlist.my, a prominent car website in Malaysia that provides used car listing. The website provides both new and used car listings and contain listings from other used car trading platform such as Carsome. To extract the data from used car listings in Carlist.my, a web scraping script has been written. The script uses Selenium with undetected chromedriver to automate browser actions, since the website being scraped implements anti-bot mechanisms. Delay was added to mimic real human behaviour such as viewing car details during browsing listings. The script goes through every listing of a page, scraping every specification listed in the car specification sections using the HTML components located using XPath. For efficiency purposes, the script will also check for duplicate listings identified by their URL, only adding new data into the Excel file.

The script starts with configurating Chrome to run in headless mode and undetected to avoid bot detection on the website.

```python
# ✅ Configure undetected-chromedriver options
options = uc.ChromeOptions()
options.add_argument("--window-size=1920,1080")
options.add_argument("--no-sandbox")
options.add_argument("--disable-dev-shm-usage")
options.add_argument("--disable-blink-features=AutomationControlled")
options.add_argument("--log-level=3")
options.add_argument(
    "--user-agent=Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/91.0.4472.124 Safari/537.36"
)

# ✅ Initialize undetected-chromedriver with the specified options
driver = uc.Chrome(options=options, headless=True)
```

Figure 17. Browser Configuration

Next the script will load the Excel file and extract the existing scraped URLs for duplicate checking during the scraping process. The script next iterates through pages, clicking on each car listing and extract the information. The extracted information will be verified with the URL loaded from the Excel file earlier to check for duplicates and dropping them.

```
# ✅ Load existing data if the file exists
if os.path.exists(EXCEL_FILE):
    df_existing = pd.read_excel(EXCEL_FILE)
    existing_links = set(df_existing["Link"].astype(str))  # Convert to set for fast lookup
else:
    df_existing = pd.DataFrame()
    existing_links = set()

cars_data = []  # List to store new extracted data
```

Figure 18. Script Loading Existing Listings

The location of information to be scraped are located and identified using the XPath that can be obtained using inspect element tool. The location is specified in the script, and a delay is added to mimic human behaviour and to wait for dynamic content to be loaded.

```
try:
    # ✅ Extract Car Price
    price_element = WebDriverWait(driver, timeout: 10).until(
        ec.presence_of_element_located(
            (By.XPATH, "//*[@id='details-gallery']/div/div/div[1]/div[1]/div[2]/div/div[1]/h3")
        )
    )
    car_details["Price"] = price_element.text.strip()
```

Figure 19. Added Delay and Specified XPath

The information scraped is then saved into an Excel file. The browser opened for scraping will be closed.

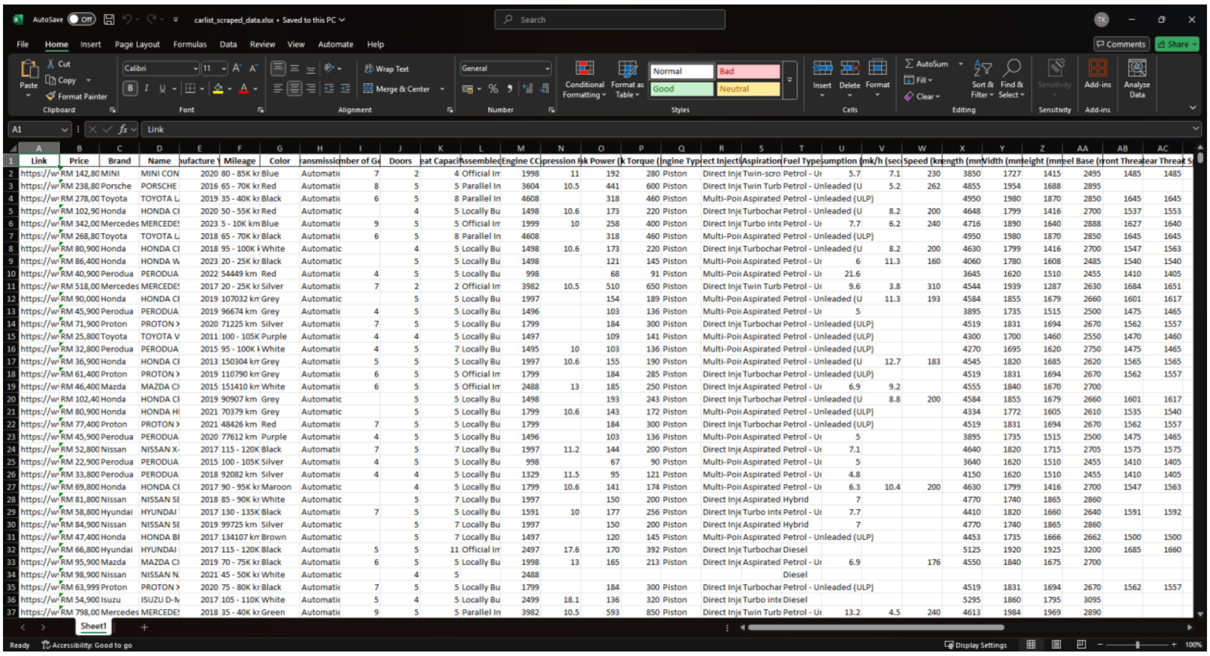


Figure 20. Excel File for Storing Scraped Data

Data Quality Report

Data Statistics

Table 4. Data Quality Report

Number of Rows	4908
Number of Columns	37
Total Missing Values	25,406

Column Name	Missing Value
Link	0
Price	0
Brand	0
Name	0
Manufacture Year	0
Mileage	0
Color	0
Listing Date	0
Transmission	0
Number of Gears	1567

Gear Type	210
Doors	0
Seat Capacity	0
Assembled	33
Engine CC	88
Peak Power (hp	57
Peak Torque (Nm)	1022
Direct Injection	1428
Aspiration	1827
Fuel Type	0
Fuel Consumption	3539
CO2 Emission	3844
Top Speed (km/h)	3195
Length (mm)	4
Width (mm)	4
Height (mm)	4
Wheel Base (mm)	4
Kerb Weight (kg)	884
Boot Space (litres)	2292
Fuel Tank (litres)	924
Front Brakes	1295
Rear Brakes	1301
Front Suspension	1197
Rear Suspension	1208
Steering Type	1044
Front Tyre	328
Rear Tyre	328
Front Rim (inches)	275
Rear Rim (inches)	273
Steering Turn	1687
Front Thread	2154
Rear Thread	2150

5.4.2 Data Preparation

Another script used to preprocess the data is written to handle missing values and dropping unnecessary columns, since there were 36 columns scraped. The columns to keep were narrowed down to 13, which are: "Link", "Price", "Brand", "Name", "Manufacture Year", "Transmission", "Doors", "Seat Capacity", "Engine CC", "Fuel Type", "Mileage", "Engine Type", and "Listing Date". Some columns values were also truncated and formatted for handling purposes. For instance, the value “Petrol - Unleaded (ULP)” in rows were truncated into “Petrol”, values of Mileage containing strings like “KM” were removed and converted in to float. Upon running the script, the duplicates were removed, invalid values are filtered out, and the cleaned dataset is saved into a new Excel to avoid making changes to the original Excel file.

```
EXCEL_FILE = "carlist_scraped_data.xlsx"
CLEANED_FILE = "cleaned_carlist_data.xlsx"
```

Figure 21. Creating a Separate Excel File for Cleaned Dataset

```
columns_to_keep = [
    "Link", "Price", "Brand", "Name", "Manufacture Year", "Transmission",
    "Doors", "Seat Capacity", "Engine CC", "Fuel Type", "Mileage",
    "Engine Type", "Listing Date"
]
```

Figure 22. Selected Columns to Include in Cleaned Dataset

```
# Remove duplicates based on the 'Link' column
df_cleaned = df_cleaned.drop_duplicates(subset=["Link"], keep="first")

# Remove invalid names
df_cleaned = remove_invalid_names(df_cleaned)
```

Figure 23. Removing Duplicates and Invalid Values

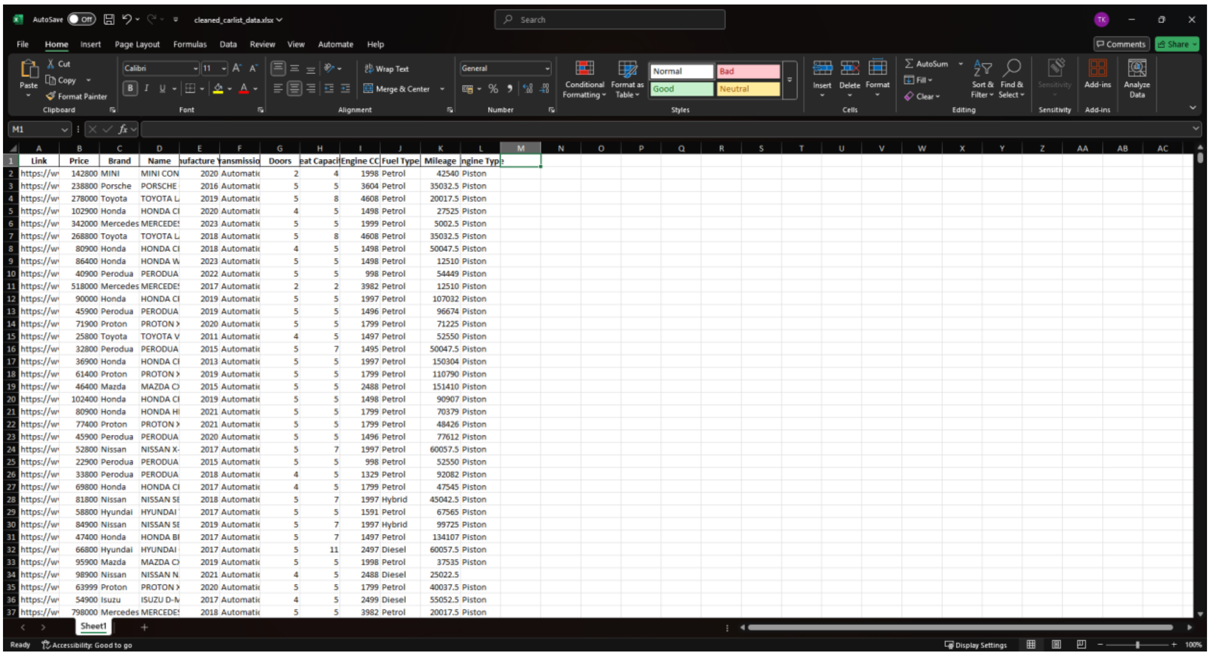


Figure 24. New Excel File for Cleaned Data

After cleaning the data, a few important features for choosing a used car were identified and extracted for preprocessing. The features include numerical features such as “Price”, “Manufacture Year”, “Mileage” and categorical features such as “Brand” and “Fuel Type”. These preprocessed features will be saved in a separate file for recommendation model building.

The missing numerical values were first handled using median strategy from SimpleImputer because the data is skewed. Median strategy allows a robust model towards outlier and skewed distribution. The skewed feature such as “Price” and “Mileage” was applied with log transformation as well to reduce skewness and improve the normality, which can improve the model performance.

```
# Impute missing values
cat_imputer = SimpleImputer(strategy="most_frequent")
df_cat_imputed_array = cat_imputer.fit_transform(mapped.values.reshape(-1, 1))
df_cat_imputed = pd.DataFrame(df_cat_imputed_array, columns=['Fuel Type'])
```

Figure 25. Data Preprocessing

Categorical features on the other hand were handled using most frequent strategy from SimpleImputer, which replace missing Fuel Type with the most common Petrol. Only few samples have missing value for Fuel Type. After that, numerical features were scaled with MinMaxScaler and applied with Domain-Specific Weights to reflect each features' importance.

The features were then standardized using StandardScaler, which is often ideal for clustering models such as K-Means. The final step of data preprocessing is saving the output in a separate file, exclusively used for modelling.

The summary statistics of the preprocessed data are also printed to check for any remaining missing values.

5.4.3 Modelling

Direct Searching

When users fill in the preference forms and click apply filter, the system will scan the dataset and return directly matching cars. The results are organized by Brand and Model for easy comparison. On the other hand, if there are no exact matches, a message "No exact match is found, please readjust your filter" will be shown.

Recommendation

Few clustering methods were chosen to test their performance in clustering the data. The methods chosen were K-Means, GMM and Hierarchical Clustering. The dataset obtained from data preprocessing will be used to develop the model, as it contains the features that are often important in selecting used cars. Before using it in the model, the dataset is first loaded and scaled features are extracted.

K-Means

The first clustering method chosen to be test is K-Means, a centroid-based algorithm. To determine a suitable number of clusters for K-Means clustering model, k numbers ranging from 2 to 10 were tested on, and the evaluation metrics such as silhouette score, CH score, DB score and inertia for each k number are printed out. Elbow Method graph is also generated for better visualization and interpretation.

```python
for k in range(2, 11):
    print(f"Testing k={k}...")
    kmeans = KMeans(n_clusters=k, init='k-means++', random_state=42, n_init=10)
    labels = kmeans.fit_predict(X_scaled)

    # Calculate evaluation metrics
    silhouette = silhouette_score(X_scaled, labels)
    ch_score = calinski_harabasz_score(X_scaled, labels)
    db_score = davies_bouldin_score(X_scaled, labels)
    inertia = kmeans.inertia_

    # Get cluster sizes
    cluster_sizes = pd.Series(labels).value_counts().sort_index().tolist()

    results.append({
        'k': k,
        'silhouette': silhouette,
        'calinski_harabasz': ch_score,
        'davies_bouldin': db_score,
        'inertia': inertia,
        'cluster_sizes': cluster_sizes
    })

    print(f"  k={k}: Silhouette: {silhouette:.4f}, CH: {ch_score:.4f}, DB: {db_score:.4f}")
```

Figure 26. Data Preprocessing

From the Elbow Method Graph, there is not a clear sharp "elbow" in the graph. Therefore, it is difficult to determine an optimal number of clusters. However, the printed evaluation metrics shows a good comparison between different k numbers. Judging from the Silhouette Score, the best performing k is 3, with a silhouette score of 0.3370, CH score at 1765.7743, DB score at 1.0340. Using CH score, K with 4 is the best performing winner, with CH score at 1808.0727, Silhouette Score at 0.3012, and DB score at 1.0099. DB score winner is the k at 10, with DB score at 0.9553, Silhouette Score at 0.2896, and CH Score at 1518.4670.

Table 5. Best K Using Different Metrics

Metric	Best K
Silhouette Score	3
CH Score	4
DB Score	10

From the result, it is seen that there is no consensus across all metrics, therefore a K number that performs good across 3 metrics will be selected. K number 4 is then selected due to its good result compared with other numbers.

However, after checking the cluster balance, it was found at k = 4, the cluster is imbalance (2056, 1014, 610, 180). Since the project objective is to recommend cars, it would be great to recommend users with recommendations that are more granular, satisfying different needs of users. Although clusters with lower K value might have better score in evaluation metrics, the recommendations may be too broad and less useful for nuanced recommendation. Therefore, a bigger cluster number consisting of balanced cluster size might perform better in this case. This led to the inspection of higher k values. K values of 6 and 7 are selected since they are granular enough to provide more specific recommendations. K values larger were not considered due to very niche clusters (eg. 1 cluster size of 76). While checking cluster size, the difference between largest and smallest cluster for K = 6 was too big (1517 vs 176). Therefore K = 7 was finally selected (1010 vs 176).

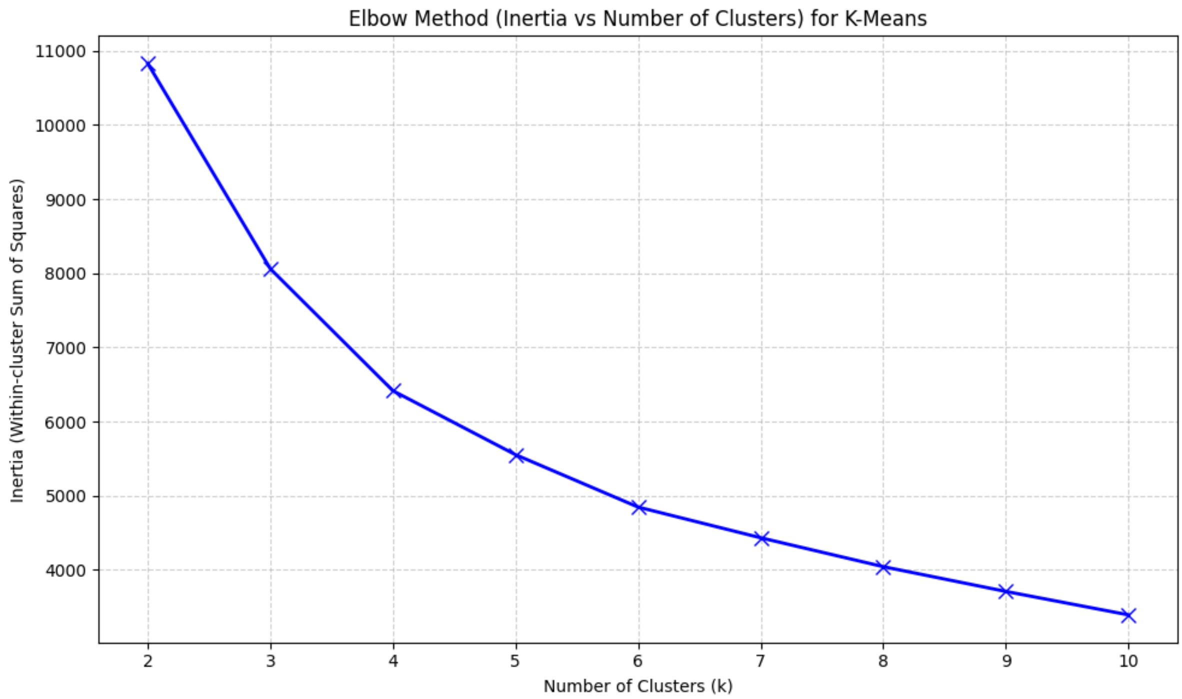


Figure 27. Elbow Method Graph Obtained

```
  k=2 → Silhouette: 0.3207, CH: 1643.1190, DB: 1.2879, Sizes: [Cluster 0: 2445, Cluster 1: 1415]
  k=3 → Silhouette: 0.3370, CH: 1765.7743, DB: 1.0340, Sizes: [Cluster 0: 2205, Cluster 1: 1423, Cluster 2: 232]
  k=4 → Silhouette: 0.3012, CH: 1808.0727, DB: 1.0099, Sizes: [Cluster 0: 180, Cluster 1: 2056, Cluster 2: 610, Cluster 3: 1014]
  k=5 → Silhouette: 0.2718, CH: 1715.7267, DB: 1.1109, Sizes: [Cluster 0: 541, Cluster 1: 1737, Cluster 2: 1126, Cluster 3: 280, Cluster 4: 176]
  k=6 → Silhouette: 0.2828, CH: 1684.5328, DB: 1.0204, Sizes: [Cluster 0: 376, Cluster 1: 285, Cluster 2: 715, Cluster 3: 1517, Cluster 4: 176, Cluster 5: 791]
  k=7 → Silhouette: 0.2628, CH: 1594.4709, DB: 1.1069, Sizes: [Cluster 0: 946, Cluster 1: 1010, Cluster 2: 176, Cluster 3: 693, Cluster 4: 483, Cluster 5: 232, Cluster 6: 320]
  k=8 → Silhouette: 0.2700, CH: 1550.9311, DB: 1.0797, Sizes: [Cluster 0: 875, Cluster 1: 213, Cluster 2: 289, Cluster 3: 160, Cluster 4: 97, Cluster 5: 963, Cluster 6: 755, C
luster 7: 508]
  k=9 → Silhouette: 0.2659, CH: 1522.2937, DB: 1.0403, Sizes: [Cluster 0: 846, Cluster 1: 156, Cluster 2: 480, Cluster 3: 227, Cluster 4: 165, Cluster 5: 741, Cluster 6: 76, C
luster 7: 335, Cluster 8: 834]
  k=10 → Silhouette: 0.2896, CH: 1518.4670, DB: 0.9553, Sizes: [Cluster 0: 453, Cluster 1: 596, Cluster 2: 156, Cluster 3: 166, Cluster 4: 226, Cluster 5: 335, Cluster 6: 777,
 Cluster 7: 73, Cluster 8: 211, Cluster 9: 867]
```

Figure 28. Evaluation of Hierarchical Clustering

GMM

The second clustering method tested is GMM. GMM is a clustering method focusing on probability of where the data point is located. Therefore, each data point has a probability of belonging to each cluster.

Similar with Kmeans, cluster numbers from 2 to 10 are tested using GMM and three different evaluation metrics, namely Bayesian Information Criterion, Akaike Information Criterion and Silhouette Score. From the result, k = 10 performs the best in BIC and AIC, while k = 2 is the best according to Silhouette Score. As mentioned before, higher k number would be better since it can provide more granular recommendations, compared to low k number that provides broad, coarse recommendations that cannot fulfill certain specific user requirements. Therefore k = 10 is selected and will be tested again with Silhouette Score, CH score and DB score. The results were unsatisfactory, as the CH score is 0.2081, CH score at 1337.2 and DB score at 1.1719. 10 clusters also suggest cluster imbalance as the smallest cluster has 7 samples only, compared to the largest cluster at 1869. Despite that, k = 10 remains the best cluster number for GMM and will be used to compare with other clustering methods to determine the best method.

```
Best k by BIC: 10
Best k by AIC: 10
Best k by Silhouette: 2
Using k = 10 for final GMM (based on lowest BIC)
```

Figure 29. AIC and BIC of GMM

```
✅ GMM Final Clustering (k=10):
  Silhouette: 0.1401
  Calinski-Harabasz: 493.3611
  Davies-Bouldin: 1.4663
  Cluster sizes:
0     191
1       7
2      12
3     167
4      63
5     410
6     947
7      69
8     125
9    1869
```

Figure 30. Evaluation and Cluster Size of GMM

Hierarchical Clustering

There 4 different linkages in hierarchical clustering, namely “single”, “average”, “complete” and “ward”. Each of these linkages will be tested from k number 2 to 10, and evaluation using Silhouette Score, CH score, DB score. The cluster size will also be reviewed to check for imbalance. A dendrogram for each linkage will also be printed out to check for the best number of k.

“Single”, “Average” and “Complete” linkage are eliminated immediately as they produce singleton cluster, which will heavily disrupt the clusters’ balance. That leaves Ward linkage the only viable option.

```
Dendrogram for single saved in models_results\Hierarchical_single
Clusters: 2, Silhouette: 0.6802, CH: 13.9950, DB: 0.2276
  Cluster sizes: [5369    1]
Clusters: 3, Silhouette: 0.5414, CH: 10.1304, DB: 0.3029
  Cluster sizes: [5368    1    1]
Clusters: 4, Silhouette: 0.5331, CH: 19.9104, DB: 0.3293
  Cluster sizes: [5365    3    1    1]
Clusters: 5, Silhouette: 0.5316, CH: 20.9843, DB: 0.3343
  Cluster sizes: [5363    3    1    1    2]
Clusters: 6, Silhouette: 0.5316, CH: 20.6365, DB: 0.3162
  Cluster sizes: [   3    2 5362    1    1    1]
Clusters: 7, Silhouette: 0.5110, CH: 17.2995, DB: 0.3142
  Cluster sizes: [5362    2    2    1    1    1    1]
Clusters: 8, Silhouette: 0.4830, CH: 15.7541, DB: 0.3172
  Cluster sizes: [5361    2    2    1    1    1    1    1]
Clusters: 9, Silhouette: 0.4814, CH: 16.6383, DB: 0.3416
  Cluster sizes: [   2 5359    2    2    1    1    1    1    1]
Clusters: 10, Silhouette: 0.4812, CH: 14.8196, DB: 0.2997
  Cluster sizes: [5359    2    2    1    1    1    1    1    1    1]
Clustering metrics plot for single saved in models_results\Hierarchical_single
```

Figure 31. Cluster size of Hierarchical (Single)

The performance of each cluster number using Ward is then reviewed to choose the best cluster number.

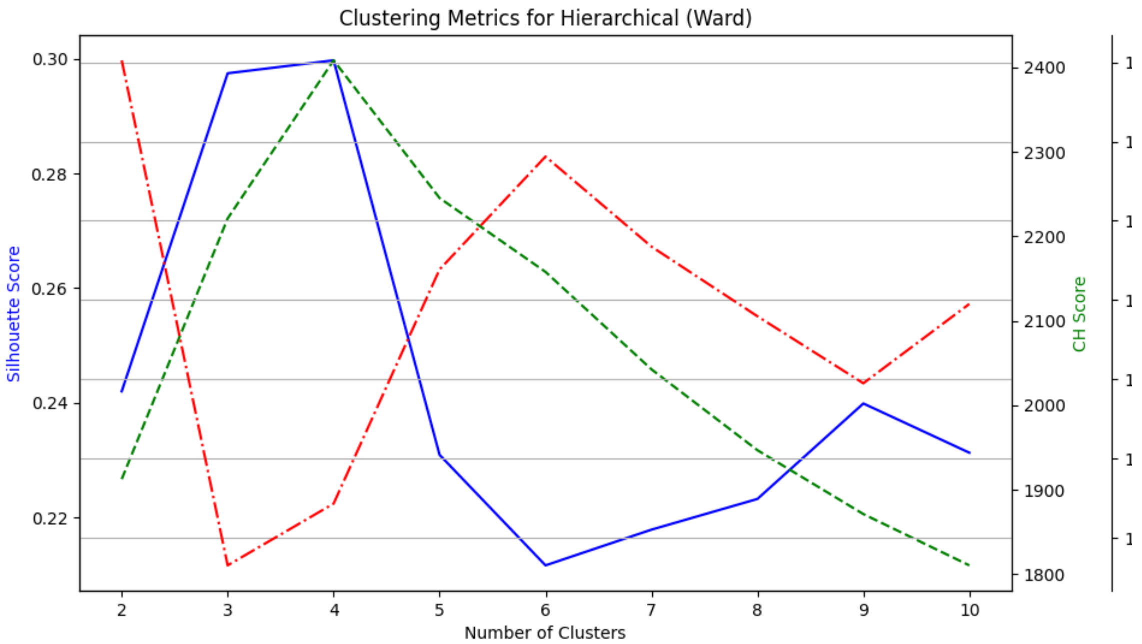


Figure 32. Silhouette Score-Number of Clusters Graph (Ward)

Evaluating using Silhouette Score and CH Score, k = 4 emerges as the best cluster number, however when looking at the cluster size, it still faces cluster imbalance (1677, 1261, 691, 231). After taking cluster balance into consideration, k = 7 was finally selected as the best cluster number as the cluster was balanced (1141, 845, 832, 691, 157, 120, 74) and has a satisfactory Silhouette Score (0.2042), CH Score, (1307.7889) and DB score (1.0824) when compared to other k numbers.

```
🖍 Testing Hierarchical Clustering with Linkage: WARD
✅ Dendrogram saved for ward
  k=2 → Silhouette: 0.2167, CH: 1214.6921, DB: 1.3058, Sizes: [Cluster 0: 2599, Cluster 1: 1261]
  k=3 → Silhouette: 0.2739, CH: 1490.2317, DB: 1.0751, Sizes: [Cluster 0: 2368, Cluster 1: 1261, Cluster 2: 231]
  k=4 → Silhouette: 0.2303, CH: 1546.1955, DB: 1.1932, Sizes: [Cluster 0: 1261, Cluster 1: 1677, Cluster 2: 231, Cluster 3: 691]
  k=5 → Silhouette: 0.2166, CH: 1389.4749, DB: 1.1591, Sizes: [Cluster 0: 1677, Cluster 1: 691, Cluster 2: 231, Cluster 3: 1141, Cluster 4: 120]
  k=6 → Silhouette: 0.1975, CH: 1322.2597, DB: 1.1624, Sizes: [Cluster 0: 231, Cluster 1: 691, Cluster 2: 832, Cluster 3: 1141, Cluster 4: 120, Cluster 5: 845]
  k=7 → Silhouette: 0.2042, CH: 1307.7889, DB: 1.0824, Sizes: [Cluster 0: 691, Cluster 1: 1141, Cluster 2: 832, Cluster 3: 157, Cluster 4: 120, Cluster 5: 845, Cluster 6: 74]
  k=8 → Silhouette: 0.2114, CH: 1300.2990, DB: 1.1119, Sizes: [Cluster 0: 832, Cluster 1: 1141, Cluster 2: 309, Cluster 3: 382, Cluster 4: 120, Cluster 5: 845, Cluster 6: 74,
Cluster 7: 157]
  k=9 → Silhouette: 0.2386, CH: 1304.8097, DB: 1.0829, Sizes: [Cluster 0: 1141, Cluster 1: 382, Cluster 2: 309, Cluster 3: 249, Cluster 4: 120, Cluster 5: 845, Cluster 6: 74,
Cluster 7: 157, Cluster 8: 583]
  k=10 → Silhouette: 0.2427, CH: 1285.7278, DB: 1.1378, Sizes: [Cluster 0: 309, Cluster 1: 382, Cluster 2: 595, Cluster 3: 249, Cluster 4: 120, Cluster 5: 845, Cluster 6: 74,
Cluster 7: 157, Cluster 8: 583, Cluster 9: 546]
```

Figure 33. Cluster Size of Hierarchical Clustering (Ward)

<u>Summary of Candidate Models</u>

Table 6. Summary of Candidate Models

Algorithm	Optimal K	Justification
K-Means	7	Granular enough, balanced clusters and good metrics
GMM	10	Best AIC and BIC amongst all K Value
Hierarchical	7 (Ward)	Balanced clusters, acceptable metrics

5.5 Dashboard

Initially, Streamlit library was used to build the dashboard for user navigation. The dashboard will implement model obtained from 4.2.4. For now, the dashboard can recommend two groups of cars, first being cars that direct matches user filter, and second being car clusters that are close to the matched cars.

At the home page of the dashboard, users can enter their filters at the left panel. Users can leave some filters blank if they have no specific requirements for certain attributes. After completing the filters, user can click the “Apply” button, and the system will show the results on the right panel.

For users who are looking for a specific car model, they can type in their car brand preferences, and there will be a new filter for car model under that car brand.

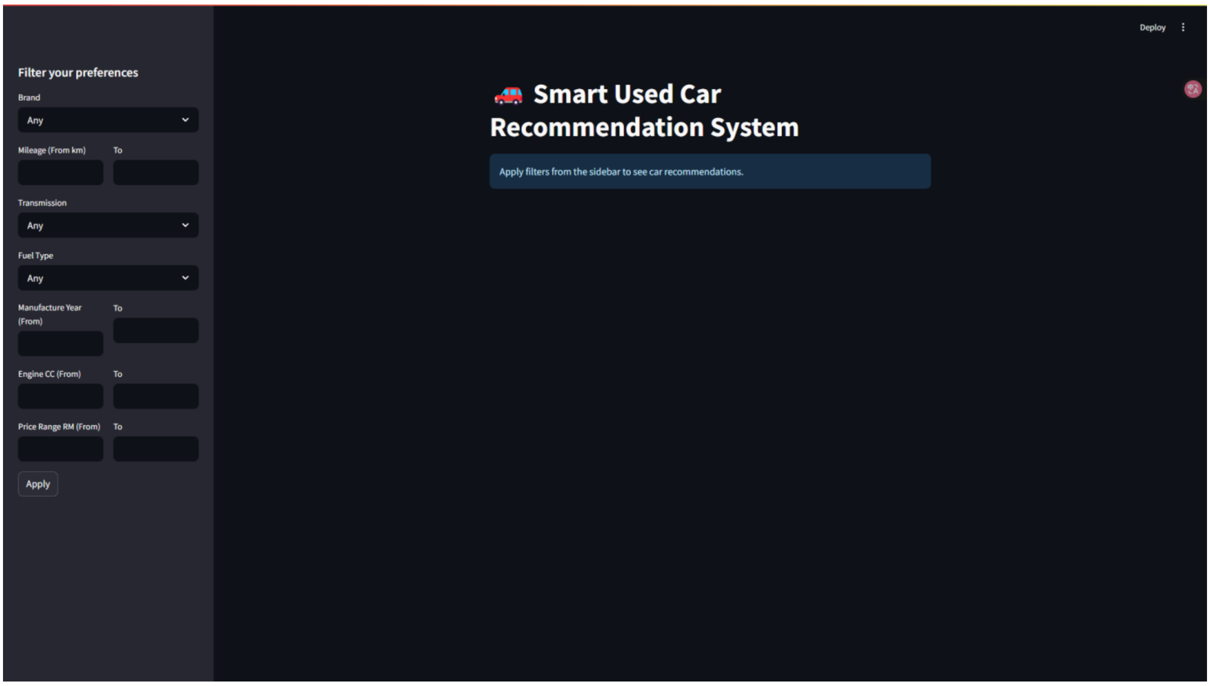


Figure 34. Home Page of the Dashboard

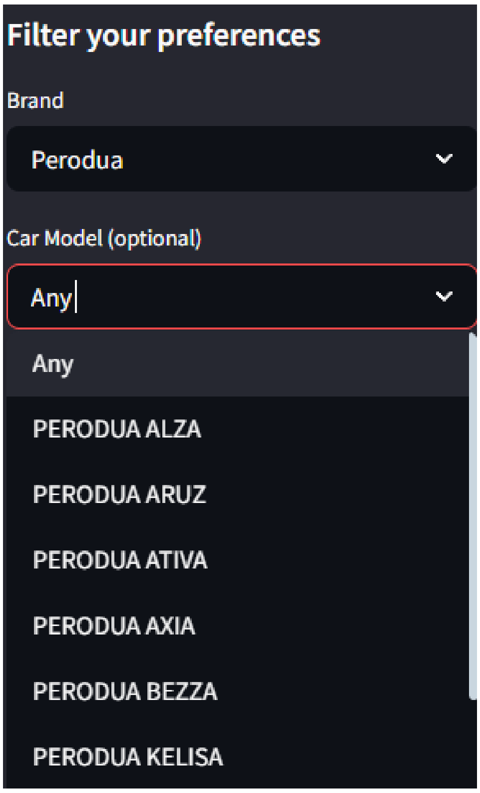


Figure 35. New Filter Option for Car Model After Completing Car Brand Selection

The first section of the recommendations shows results that directly matches user filters. The results are car models differentiated by their manufactured years, and when user click on each result, they can view the detailed specifications of the car and listing URLs. In future we plan to add comment sections under the listing URLs section.

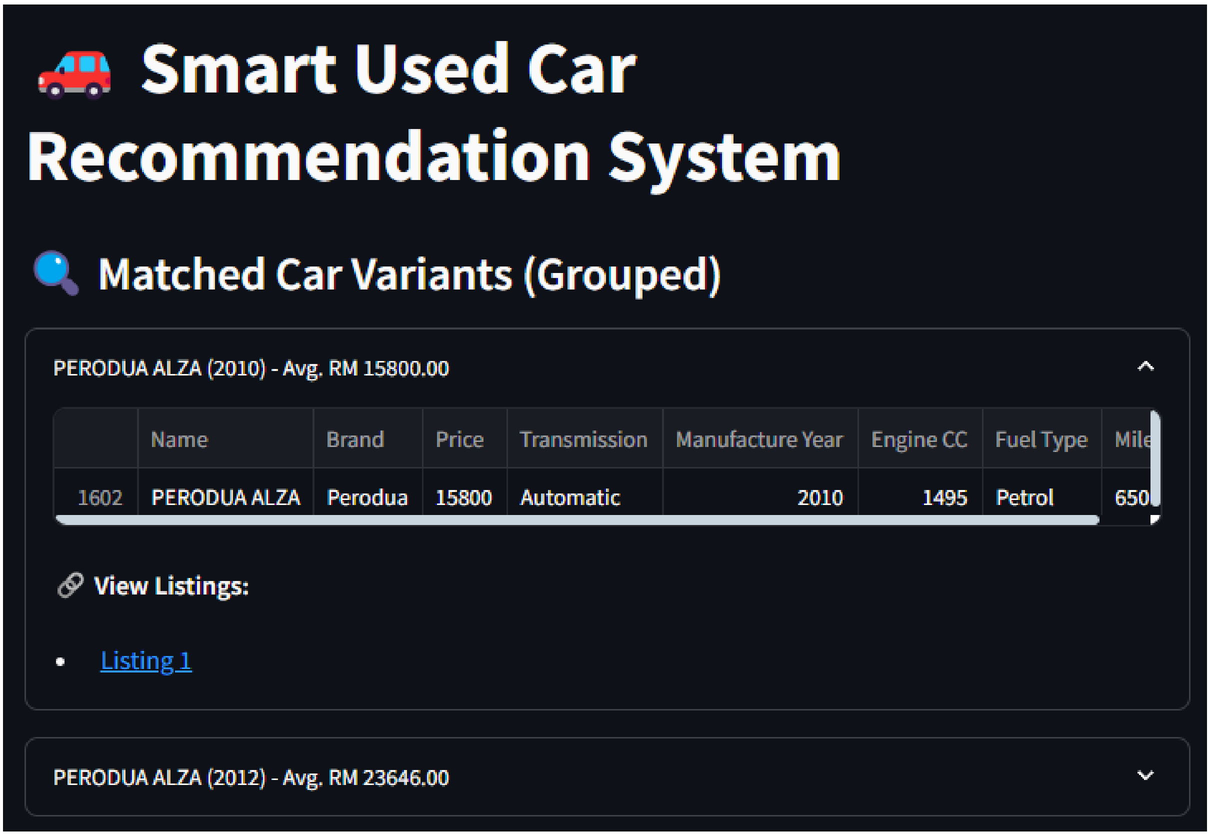


Figure 36. First Section of Recommendation

Below the first section of recommendations, there is a second section that shows car clusters that are close to the directly matched results. The purpose of this section is to recommend similar cars that users might like. Similar with the first section, the second section also includes detailed specifications. In future, we plan to enhance this section by adding explanations why users might like these second sections' recommendations. For example, similar attributes like Engine CC.

Figure 37. Second Section of Recommendations

However, due to limited customization options offered by Streamlit, the platform was later switched to Django, with the aim of designing interface that is more appealing and interactive. Django uses Model-Template-Controller structure, which is completely different from Streamlit, which only requires handle all logic and presentation in a python file. Django has View and Templates which are used to handle presentation and business logic separately and features a lot more in-built functionality than Streamlit.

A prototype was first created rapidly on Django by creating a HTML Template that resembles the original Streamlit version. The business logic is then migrated into a syntax that suits Django in the View. After checking to ensure that the application works as intended, the restructuring of the application interface design begins.

The design starts with adding a homepage with a car background and a catchy tagline for visual stimulation. Animations were also added to further enhance the applications' appearance. A button that redirects user to the recommendation page is then added.

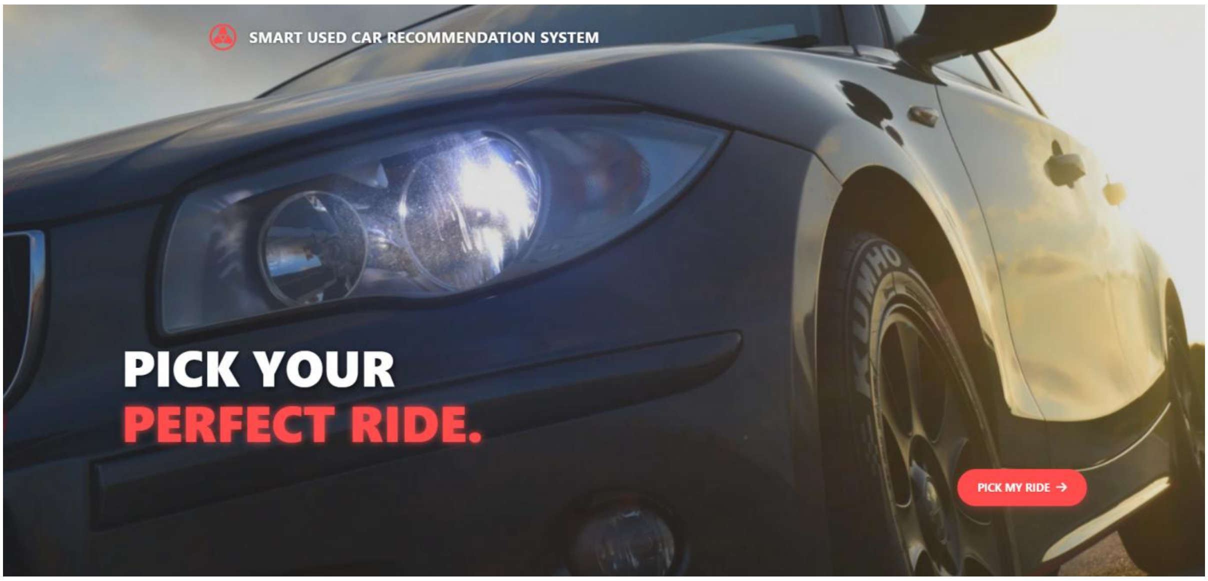


Figure 38. Home Page

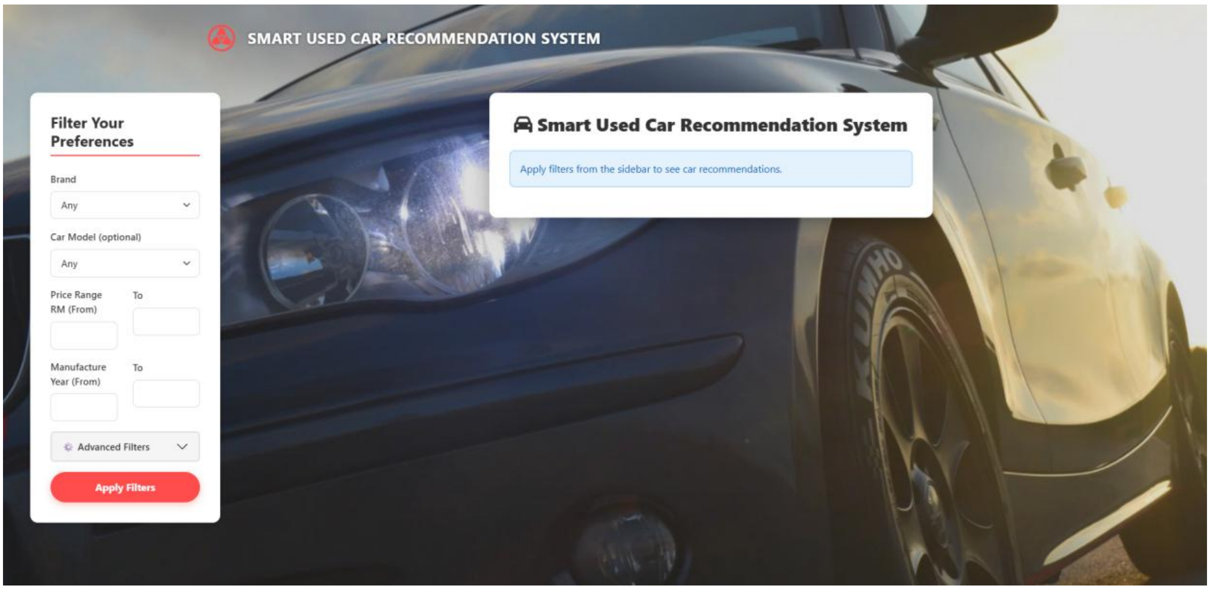


Figure 39. Recommendation Page

CHAPTER 6

System Evaluation and Discussion

In this chapter, evaluation will be conducted on the models to select one best performing models using evaluation metrics. The best selected method will then be deployed to integrate with the dashboard. Discussion of challenges faced during the project are also going to be presented, showing how these challenges were addressed. Finally, the fulfilment of the project objective will also be assessed.

6.1 System Evaluation

During the modelling phase, three models, namely Kmeans, Hierarchical, GMM were trained. To assess the clustering quality and the best number of clusters each model offers, evaluation metrics such as Silhouette Score, Davies-Bouldin Index, Calinski-Harabasz Index and more. Using evaluation metrics, an optimal K number will be selected from each model and then compared with the best K numbers from other models to determine the final optimal K number and the final optimal clustering method.

So far, the selected k number of each clustering methods are: KMeans with 7, GMM with 10 and Hierarchical Clustering (Ward) with 7. Table below shows the comparison of metrics across all three algorithms:

Table 7. Best K Value Chosen from Each Method

Metric	Kmeans (7)	Hierarchical (Ward, 7)	GMM (10)
Silhouette Score	0.2628	0.2042	0.198
CH Score	1594.47	1307.79	1337.2
DB Score	1.107	1.0824	1.1719
Cluster Size	(1010, 946, 693, 483, 320, 232, 176)	(1141, 845, 832, 691, 157, 120, 74)	(1869, 947, 410, 191, 167, 125, 69, 63, 12, 7)

Largest Cluster	1010 (25.9% of all data)	1141 (29.3% of all data)	1869 (48.0% of all data)
Smallest Cluster	176 (4.5% of data of all data)	74 (1.9% of all data)	7 (0.02% of all data)

GMM is immediately eliminated due to the presence of extreme clusters. For instance, the smallest cluster has only 7 samples, while the largest cluster has 1869 samples, accounting for 0.02% and 48% of the full data respectively. Despite being the best optimal K value from GMM, it lacks the suitability for the project's business objective. This leaves K-Means and Hierarhical Clustering as the remaining models to be compared.

The comparison shows that Hierarhical (Ward) has a slightly better DB score (1.0824 vs 1.107), indicating better compactness. However, K-Means has significantly outperformed Hierarchical (Ward) in Silhouette Score and CH Score, marking increase of 28.7% and 21.9% respectively. The minimum cluster size of K-Means (4.5% of all data) is also better than the minimum cluster size of Hierarchical Ward (1.9% of all data), indicating that KMeans has a better cluster balance.

Therefore, the final chosen model for development is KMeans with cluster number 7. The justifications were that K-Means with 7 has the highest overall clustering evaluation metrics (Best Silhouette Score and CH Score). The clusters were also balanced and granular enough, which allows diverse recommendations. For instance, the 7 distinct clusters identified represents car groupings that are not too niche or broad.

6.2 Deployment

The data is clustered with final selected K = 7 using KMeans, and a description for each cluster is generated to provide summaries of each cluster based on the centroid feature values (Price range, Mileage range, etc.). To enhance user engagement in the dashboard section, appealing marketing descriptions were added.

```python
# Custom descriptions based on K-Means clusters
if cluster_id == 0:
    description = f"Luxury at its finest. Mainly {avg_year} {dominant_brand}s with {dominant_fuel} engines, averaging RM{avg_price:,}, low mileage ({avg_mileage:,} km)."
elif cluster_id == 1:
    description = f"Premium {dominant_brand}s, {dominant_fuel}-powered, around {avg_year}, averaging RM{avg_price:,}, moderate mileage ({avg_mileage:,} km)."
elif cluster_id == 2:
    description = f"Affordable {dominant_brand}s, {dominant_fuel} engines, around {avg_year}, RM{avg_price:,}, higher mileage ({avg_mileage:,} km)."
elif cluster_id == 3:
    description = f"Efficient mid-range {dominant_brand}s, {dominant_fuel}, around {avg_year}, RM{avg_price:,}, moderate mileage."
elif cluster_id == 4:
    description = f"Next-gen EVs, around {avg_year}, very low mileage ({avg_mileage:,} km), sustainable modern choice."
elif cluster_id == 5:
    description = f"Fresh, affordable rides ({avg_year}), mostly {dominant_brand}s, RM{avg_price:,}, moderate mileage."
elif cluster_id == 6:
    description = f"Older, rugged {dominant_brand}s, {avg_year}, high mileage ({avg_mileage:,} km), budget-friendly option."
```

Figure 40. Cluster Description

The trained K-Means model was then saved into a .pkl file for integration into the dashboard. This allows the dashboard to recommend cars immediately based on user preference. The method also allows scalability, where future model updates can be used by the dashboard immediately.

6.3 Project Challenges

Throughout the development of this project, several challenges were encountered, where each played an essential part in shaping the system's performance and appearance.

The first challenge was the limited data size. By obtaining data through web scraping, not only it was time consuming, the data collected were also insufficient. For example, it takes 3 hours to scrape 10 pages of data, with each page containing 25 car listings, for a total of 250 car listings. Considering the possibility of missing values in the car listings, the final data obtained may be even less. As a result, the limited data size may affect the model's ability to cluster, affecting the clustering quality.

Another challenge was data quality. For example, the data initially captured contained 4,908 samples, which were reduced to 3,861 after data cleaning, a total reduction of nearly 1,000 samples. This worsens the already limited data size.

The next challenge was transitioning from Streamlit to Django. Creating a Django dashboard and integrating it with the model was much more complex than with Streamlit and required more time and effort. While the Model-Template-Controller structure of Django provides more customizability and separation of concern, it also introduces a steeper learning curve.

6.4 Objective Evaluation

To identify a suitable machine learning algorithm in providing used car recommendations

The primary objective of identifying a suitable machine learnig algorithm that provide car recommendations was achieved. It was done through deploying various algorithms, specifically clustering such as K-Means, GMM and Hierarhical to group cars based on features. K-Means was finally selected as the model to be developed due to its alignment with business objectives. The evaluation metrics has also confirmed the algorithms' ability to cluster cars into suitable groups.

To develop and train a model that can give relevant recommendations to user

The selected K-Means with optimal K = 7 was trained on the preprocessed data set, assigning every car to the clusters. Each cluster was then assigned with technical descriptions and consumer-friendly, marketing-focused descriptions to provide a detailed overview of the cluster characteristics.

To design and implement a dashboard interface that allows users to enter their inputs of desired used car specifications

A dashboard was designed and implemented using Django, allowing users to enter their preferences. The dashboard then provides recommendations group by same car brands and similar clusters.

CHAPTER 7

Conclusion and Recommendation

This chapter provides a comprehensive summary of the project outcome, highlighting its contribution to the used car market in Malaysia. Recommendations of approaches that can be implement in future will also be provided in this chapter, with the hope of further enhancing the project features.

7.1 Conclusion

In conclusion, the developed used car recommendation system can help to address the problem of lacking a similar system catered to Malaysians who are in search for used car recommendations, by providing them with a few relevant recommendations and improving their purchasing decision. A series of literature review has been conducted to uncover relevant algorithms that has been applied in the field of car recommendation system. Existing online tools such as CarChooser and CarBase has also been reviewed, providing useful insights and feature ideas to be added in our project.

The existing tools act as a foundation for understanding of how the design of the system should be. The project did not only cover the features provided by them, but also introduced recommendation features, as it was not included in the tools, where they only use filter-based model. CRISP-DM is the methodology to follow during the development of this project, as it allows flexibility for developers to revisit any phases, should there be any changes need to be implemented.

K-Means was finally chosen as the best algorithm to cluster the data, demonstrating its ability to cluster cars. The dashboard created also allows user to input their preferences, obtaining relevant car recommendations.

7.2 Recommendation

To enhance and refine the current system in the future, several recommendations can be made.

First recommendation is to increase the sample size. Our dataset currently consists of nearly 4,000 samples, after data cleaning. To improve the model's clustering ability and accuracy, the sample size should be increased in the future. For example, data can be scraped on different machines, each processing a different car listing site to ensure there is no duplication, and the scraping process is fast. By incorporating a larger dataset, the model can potentially spot hidden patterns within the dataset, increasing the robustness of the clustering model. From user perspective, they can also access listings from various renowned car listing sites, which reduce the hassle of searching listings on different sites manually.

Another recommendation is to incorporate Natural Language Processing inputs. Considering that Artificial Intelligence is currently a major trend, from a marketing standpoint, a Natural Language Processing based used car recommendation system could be a marketing advantage, offering innovative, user-friendly experience. With Natural Language Processing, users can express their preferences in everyday language, and the system interprets the input to provide relevant recommendations. Natural Language Processing also provides conversational interaction, mimicking a real used car sales agent

REFERENCES

[1] “Passenger Cars - Malaysia | Statista Market Forecast,” *Statista*. https://www.statista.com/outlook/mmo/passenger-cars/malaysia

[2] G. of Malaysia, “Vehicle Registration Transactions: Cars | data.gov.my,” *data.gov.my*. https://data.gov.my/data-catalogue/registration_transactions_car

[3] “Malaysia overtakes Thailand as ASEAN’s second-biggest auto market,” *Nikkei Asia*. https://asia.nikkei.com/Business/Automobiles/Malaysia-overtakes-Thailand-as-ASEAN-s-second-biggest-auto-market

[4] N. Daim, “36.3 million vehicles in Malaysia | New Straits Times,” *NST Online*, Dec. 06, 2023. https://www.nst.com.my/news/nation/2023/12/987062/363-million-vehicles-malaysia

[5] “Malaysia Used Car Market Size, Share, Industry Analysis,” *www.kenresearch.com*. https://www.kenresearch.com/industry-reports/malaysia-used-car-market

[6] “Malaysia Used Car Market Size, Share, Trends | Industry Outlook (2022 - 27),” *www.mordorintelligence.com*.https://www.mordorintelligence.com/industry-reports/malaysia-used-car-market

[7] D. Tan, “2024 Malaysia car sales data by brand vs 2023 – take a look at last year’s biggest winners and losers,” *Paul Tan’s Automotive News*, Feb. 06, 2025. https://paultan.org/2025/02/06/2024-malaysia-car-sales-data-by-brand-vs-2023-take-a-look-at-last-years-biggest-winners-and-losers/

[8] B. Scheibehenne, R. Greifeneder, and P. M. Todd, “Can There Ever Be Too Many Options? A Meta-Analytic Review of Choice Overload,” *Journal of Consumer Research*, vol. 37, no. 3, pp. 409–425, Oct. 2010, doi: https://doi.org/10.1086/651235.

[9] S. S. Iyengar and M. R. Lepper, “When choice is demotivating: Can one desire too much of a good thing?,” *Journal of Personality and Social Psychology*, vol. 79, no. 6, pp. 995–1006, 2000, doi: https://doi.org/10.1037/0022-3514.79.6.995.

[10] R. Alabduljabbar, M. Alghamdi, and H. Alshamlan, “Personalized Car Recommendations Using Knowledge-Based Methods,” *2023 Intelligent Methods, Systems, and Applications (IMSA)*, pp. 539–544, Jul. 2023, doi: https://doi.org/10.1109/imsa58542.2023.10217406.

[11] M. S. Karthikeyan, A. A. Blessie, J. R. Francis Raj, R. S. Krishnan, K. Haribabu, and S. Balamurugan, “Multi-Modal Deep Learning for User-Centric Car Recommendations,” *2025 International Conference on Multi-Agent Systems for Collaborative Intelligence (ICMSCI)*, pp. 1765–1772, Jan. 2025, doi: https://doi.org/10.1109/icmsci62561.2025.10894027.

[12] H. Mahgfuri, R. D. Perdanakusuma, F. Basfianto, A. A. Sukmandhani, and I. P. Saputro, “A Recommendation System for Buying a used Car using the Promethee Method,” *2022 International Conference on Sustainable Islamic Business and Finance (SIBF)*, pp. 209–213, Oct. 2022, doi: https://doi.org/10.1109/sibf56821.2022.9939732.

[13] S. Pawar, A. Pawar, P. Pawar, and Jayashri Bagade, “Car Recommendation System,” pp. 1–6, Oct. 2024, doi: https://doi.org/10.1109/icisaa62385.2024.10828698.

[14] “carwow | Car buying with you in the driving seat,” *carwow.co.uk*, 2013. https://www.carwow.co.uk/

[15] “Car Chooser,” *CarExpert*, 2020. https://www.carexpert.com.au/car-chooser

[16] “Malaysia’s #1 Car Buyer’s Guide,” *www.carbase.my*. https://www.carbase.my/

[17] N. Hotz, “What Is CRISP DM?,” *Data Science Project Management*, Dec. 09, 2024. https://www.datascience-pm.com/crisp-dm-2/

[18] C.-D. Wang, Z.-H. Deng, J.-H. Lai, and P. S. Yu, “Serendipitous Recommendation in E-Commerce Using Innovator-Based Collaborative Filtering,” IEEE Transactions on Cybernetics, vol. 49, no. 7, pp. 2678–2692, Jul. 2019, doi: https://doi.org/10.1109/tcyb.2018.2841924.

[19] S. Chumbar, “The CRISP-DM Process: A Comprehensive Guide,” *Medium*, Sep. 24, 2023. https://medium.com/@shawn.chumbar/the-crisp-dm-process-a-comprehensive-guide-4d893aecb151

POSTER

DESIGN AND DEVELOPMENT OF A SMART USED CAR RECOMMENDATION SYSTEM

Project Developer: Te Kay Hoe
Project Supervisor :Dr Abdulkarim Kanaan Jebna

PROBLEM STATEMENT

- Choice overload overwhelming users
- Lack of recommendation system tailored to Malaysian used car consumers

PROJECT SCOPE

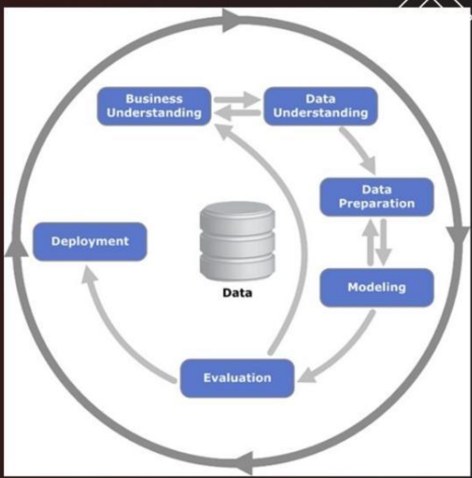
- Develop a dashboard for Malaysian used car consumers to look for recommendations

PROJECT OBJECTIVES

- Identify a suitable algorithm for the system
- Develop and train a model that can give relevant recommendations
- Design and implement a dashboard for user navigation
- Implement user engagement features to create a community-driven interactive platform

METHODOLOGY

- CRISP-DM



CONCLUSION

- The proposed dashboard aims to improve Malaysian used car consumers' purchasing decision by providing personalized recommendations

UTAR
UNIVERSITI TUNKU ABDUL RAHMAN
Faculty of Information and Communication Technology
Bachelor of Business Information Systems