Proceeding of the 18th INFATS & 27th SAE-C/STC, 2024, pp1-9
2024.106

Decision Tree Analysis on Japanese Automotive Risk Data

Ayuka SHINAGAWA[1], Xiaodong FENG [2], Kun ZHANG[3]

[1]*Department of Information and Management Systems Engineering, Nagaoka University of Technology, Nagaoka, Japan,940-2188*

[2]*Department of Information Science and Control Engineering, Nagaoka University of Technology, Nagaoka, Japan,940-2188*

[3]*Department of System Safety Engineering, Nagaoka University of Technology, Nagaoka, Japan, 940-2188*

Email: s201038@stn.nagaokaut.ac.jp

Abstract:

Background: The increasing number of automotive accidents and defects pose significant risks to road safety and manufacturer reliability. This study integrates and analyzes three major types of automotive risk data—accident/fire, defect, and recall information—using the IBM SPSS Modeler. Understanding patterns in these risks is crucial for improving vehicle safety and minimizing accidents.

Objective: To identify key factors contributing to vehicle-related accidents and defects across major automotive manufacturers, providing a comprehensive understanding that can inform risk management and safety improvements.

Methods and Materials: The study utilized a large dataset of approximately 200,000 records from 2015 to 2022. This dataset was filtered to include five major manufacturers (Toyota, Honda, Nissan, Suzuki, and Mitsubishi) and two vehicle types (passenger and light passenger cars). A decision tree analysis was conducted to explore risk factors, including the first registration year, mileage, and defective part codes.

Results: The analysis revealed significant differences in accident and defect trends by manufacturer. For some brands, higher mileage was linked to an increased likelihood of accidents or defects, while for others, the first registration year was a significant factor. The decision tree analysis provided insights into the variation of risk patterns across manufacturers.

Conclusions: The study found that accident and defect trends are influenced by a variety of factors, which vary by manufacturer. While the current analysis was limited to a smaller sample size, future research will involve larger datasets, additional analytical techniques like Bayesian network analysis, and text data analysis to improve understanding and management of automotive risks.

Keywords: Automotive Risk, Accident Analysis, SPSS Modeler, Decision Tree Analysis, Data Integration.

1 Introduction

With the continuous development of the global automobile industry, automobiles have become an indispensable means of transport in modern society. As an important automobile manufacturing and consumption market globally, Japan's domestic car ownership has been at a high level. According to the latest data, the number of domestic vehicles in Japan reached about 82 million units in 2022 [1]. However, as the number of vehicles increases, automobile safety issues are becoming more prominent, especially recalls due to mechanical failures. In 2022 alone, the number of recalls in Japan reached 383, involving approximately 4.6 million vehicles [2]. This phenomenon has a far-reaching impact on car users and puts enormous economic pressure and reputational risk on companies.

In recent years, the Ministry of Land, Infrastructure, Transport and Tourism (MLIT) has been actively promoting a policy of information disclosure, providing society with a large amount of data related to automobile safety, including automobile accidents and fires, mechanical failures, and recalls. Through this data, consumers can better understand the safety status of their own vehicles, and companies can make necessary improvements accordingly. However, despite the many measures taken by governments and companies, the frequency of car breakdowns has not been significantly reduced. This means that existing data analytics methods and tools are still inadequate in utilising this information to identify and prevent vehicle risks.

With the rise of big data analytics, especially the increasing maturity of the application of methods such as machine learning and decision trees in classification and prediction tasks, how to maximise the use of existing data resources through more advanced technological means has become an important topic of current research [3-5]. In addition, most of the existing research in Japan is limited to analysing a single type of data, failing to integrate accident, fire, defect, and recall data to assess vehicle risk as a whole. This isolated approach to analyses limits researchers and policymakers from gaining a comprehensive understanding of the root causes of the problem. To this end, this study proposes a new analytical framework to dig deeper into the patterns and laws behind vehicle risks through data integration and machine learning techniques. This study aims to perform a correlation analysis of three major databases of automotive risk information (accident and fire, defect, and recall) made public by the Ministry of Land, Infrastructure, Transport and Tourism (MLIT) using IBM SPSS Modeler. By combining these data and performing a decision tree analysis, this study expects to identify the key factors affecting vehicle risk and gain knowledge and countermeasures to reduce the occurrence of automotive accidents and defects.

2 Method and Material

2.1 Research Overview

This study utilized IBM SPSS Modeler as the primary data analysis tool, specifically for classification and prediction using decision tree algorithms. IBM SPSS Modeler is a platform widely used in big data analysis, data mining, and machine learning, known for its intuitive user interface and powerful data processing capabilities, making it popular in both academic and industrial fields. AlKheder et al. effectively utilized SPSS Modeler for machine learning analysis, showcasing its robust application in the classification and prediction of traffic accidents data [6]. Aiash et al. utilized SPSS Modeler to analyze pedestrian crash injury risk factors in Barcelona, effectively demonstrating the tool's capability in identifying key variables and patterns associated with pedestrian accidents. Their analysis leveraged decision tree algorithms to uncover significant insights into the factors influencing injury severity, highlighting SPSS Modeler's utility in transportation safety research and risk prediction [7]. SPSS Modeler supports various machine learning algorithms and can quickly process large datasets. Through this tool, the study was able to easily integrate different data sources and use decision tree algorithms to classify and predict automotive risks.

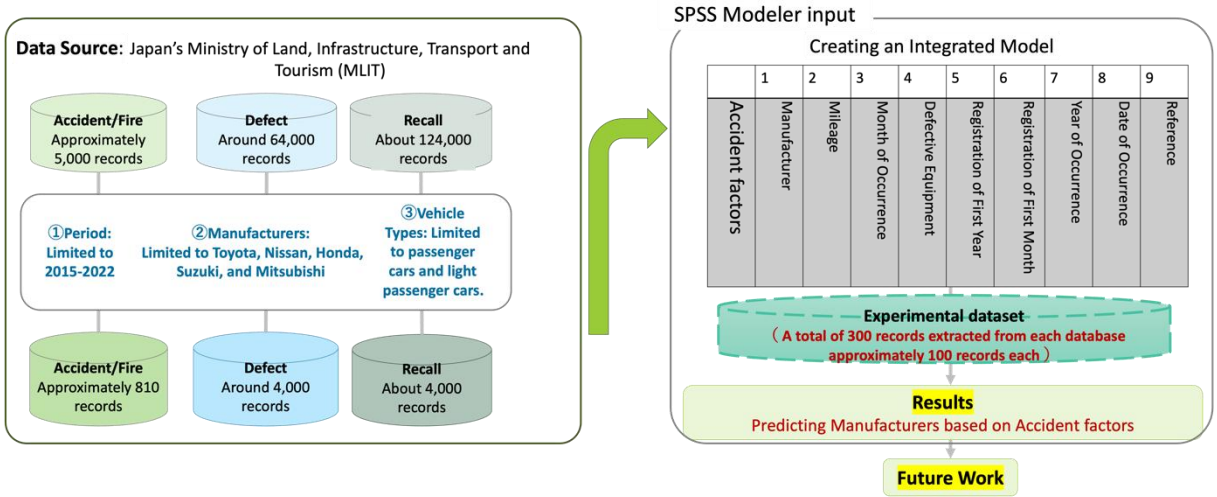

Figure 1 overview of the research on decision tree analysis on automotive risk data

The application of IBM SPSS Modeler in this study primarily involved data cleaning, feature selection, model construction, and results visualization. As a visual modeling platform, it allows researchers to intuitively build and adjust machine learning models without writing code. Therefore, the use of SPSS Modeler significantly simplified the data analysis process while ensuring the reliability and efficiency of the analysis. By utilizing SPSS Modeler's capabilities, efficient data

preprocessing, model building, evaluation, interpretation, and application are possible, enabling effective predictive analysis based on the product characteristics of different vehicle manufacturers.

Decision tree is a machine learning method widely used for classification and prediction by hierarchically classifying datasets from which key decision rules are extracted [8]. Firstly, three independent datasets are integrated into one unified dataset, followed by decision tree analysis to identify the key factors affecting vehicle risk. Through this analysis method, the relationship between different variables can be visualised, thus providing a basis for future vehicle risk management. Figure 1 shows an overview of the research.

2.2 Data collection

These datasets, originally totalling around 200,000 records, cover details on accidents, defects, and recalls. For focused analysis, the data was filtered to include cases from 2015 to 2022, limited to five major automakers (Toyota, Honda, Nissan, Suzuki, Mitsubishi) and two vehicle types (passenger and light passenger cars). After filtering, the final dataset consists of approximately 810 accident and fire cases, 4,000 defect cases, and 4,000 recall cases. The collected data underwent preliminary cleaning to remove missing values and duplicate records, followed by standardization to ensure consistency across the databases. The numerical data were also discretized, and derived variables were created, such as classifying accident occurrences by mileage.

2.3 Creating data integration models

First, the items for the integrated model were organized, and 9 items were selected: 1_ Manufacturer, 2_Mileage, 3_Month of Occurrence, 4_Defective Equipment, 5_Registration of First Year, 6_Registration of First Month, 7_Year of Occurrence, 8_Date of Occurrence and 9_Reference. Then, 100 records each were randomly selected from the accident/fire, defect, and recall database, totalling 300 records, to create the experimental dataset. The details of the experimental dataset are shown in Figure 2.

Manufacturer	Count	%
Toyota	106	35.3
Honda	43	14.3
Nissan	88	29.3
Suzuki	35	11.7
Mitsubishi	28	9.3

Year of First Registration	Count	%
-1999	24	8.0
2000-2004	51	17.0
2005-2009	65	21.7
2010-2014	84	28.0
2015-2019	65	21.7
2020-	11	3.7

Month of First Registration	Count	%
1	32	10.7
2	27	9.0
3	38	12.7
4	19	6.3
5	19	6.3
6	30	10.0
7	19	6.3
8	24	8.0
9	27	9.0
10	17	5.7
11	19	6.3
12	29	9.7

Mileage	Count	%
0-19999	31	15.5
20000-39999	29	14.5
40000-59999	33	16.5
60000-79999	26	13.0
80000-99999	17	8.5
100000-119999	19	9.5
120000-139999	13	6.5
140000-159999	13	6.5
160000-179999	6	3.0
180000-199999	5	2.5
200000-219999	2	1.0
220000-	6	3.0

Year of Occurrence	Count	%
2015	15	7.5
2016	19	9.5
2017	36	18
2018	39	19.5
2019	41	20.5
2020	36	18
2021	14	7
2022	0	0

Month of Occurrence	Count	%
1	21	10.5
2	20	10.0
3	19	9.5
4	14	7.0
5	15	7.5
6	13	6.5
7	18	9.0
8	15	7.5
9	22	11.0
10	12	6.0
11	14	7.0
12	17	8.5

Date of Occurrence	Count	%
1-5	71	35.5
6-10	21	10.5
11-15	27	13.5
16-20	25	12.5
21-25	23	11.5
26-31	33	16.5

Defective Equipment	Count	%
1_Engine	50	16.7
2_Transmission unit	24	8.0
3_Travelling unit	2	0.7
4_Control unit	8	2.7
5_Brake unit	18	6.0
6_Buffer unit	3	1.0
7_Fuel unit	7	2.3
8_Frame Body	13	4.3
9_Safety Lamp	13	4.3
10_Riding unit	61	20.3
11_Electrical unit	8	2.7
12_Exhaust gas and noise	7	2.3
13_Electric motor	5	1.7
14_Other unit	81	27.0

Reference	Count	%
1_Accident and Fire Data	100	33.3
2_Defect Data	100	33.3
3_Recall Data	100	33.3

Figure 2 Integrated data for experiments

2.4 Decision Tree Analysis

Decision tree analysis is a supervised learning technique used for classification and regression. It builds a tree-like structure where internal nodes represent feature tests, branches represent test outcomes, and leaf nodes represent predictions [9]. Decision trees are used in finance for credit scoring, healthcare for diagnostics, and marketing for customer segmentation. Different types of decision tree algorithms exist, each with its own characteristics and splitting criteria, such as CART (Classification and Regression Tree), C4.5/C5.0, and CHAID (Chi-Square Automatic Interaction Detector). These algorithms use different splitting criteria:

- Information Gain: Measures the reduction in entropy before and after a split.
- Gini Index: Assesses node impurity, with smaller values indicating more homogeneous nodes.
- Gain Ratio: A normalized form of information gain to prevent bias towards features with more categories.

In this study, the C5.0 and CHAID decision tree algorithm was used, which splits the data based on information gain and prunes the tree to avoid overfitting.

2.5 Decision Tree Analysis Using SPSS Modeler

SPSS Modeler provides a straightforward interface for building decision tree models for both classification and regression tasks. There have been many scholars who have utilized SPSS Modeler for decision tree analysis [9-13]. In this study, we aim to predict the manufacturer of vehicles that have been involved in accidents or defects based on such as the registration of first year, defective equipment, and mileage. The goal is to compare the product characteristics of vehicles from five major manufacturers by using decision tree analysis with SPSS Modeler, allowing for an efficient and visual understanding of each manufacturer's product features.

2.5.1 Data Preparation

Data preprocessing is the first step in the analysis to ensure data quality and consistency. For the three independent datasets (accident/fire, defect, recall) in this study, the preprocessing steps include:

- Handling Missing Values: Records with many missing values were removed. For cases with minor missing data, imputation methods were applied.
- Duplicate Removal: To ensure data independence and accuracy, duplicate records were identified and removed.
- Data Discretization: Numerical data such as mileage were discretised into categorical ranges for classification analysis.
- Derived Variables: To enhance the accuracy of the analysis, derived variables were created, such as categorising accident occurrences based on mileage.

To ensure model reliability and prevent overfitting, the dataset was split into 80% for training and 20% for testing, allowing for both model training and validation. Feature selection involved selecting nine key input variables based on their significance, including manufacturer, first registration year, mileage, defective part, and others. These variables form the foundation for the model's decision-making process, enabling accurate classification and prediction of outcomes.

2.5.2 Selecting a Decision Tree Algorithm

SPSS Modeler supports multiple decision tree algorithms, and the most suitable one can be selected based on the prediction objectives [14]:

- **C5.0**: A fast and accurate algorithm for classification tasks, offering features such as boosting to enhance performance [15].
- **CHAID (Chi-Square Automatic Interaction Detector)**: Splits nodes based on chi-square statistics, which is particularly useful for categorical predictors [16].
- **C&R Tree (Classification and Regression Tree)**: Applicable to both classification and regression, using criteria such as the Gini index or variance reduction for splitting nodes [17].

2.5.3 Model Building

- **Node Selection:** Select the appropriate decision tree node (C5.0, CHAID) from the modeling palette.
- **Specify Target and Input Variables:** Set the target variable as the car name and the input variables as the relevant attributes like first registration year.

- **Configure Model Parameters:** Adjust model parameters such as the maximum tree depth, minimum cases per node, and pruning options to balance model accuracy and complexity. The overview of decision tree model processing in the SPSS Modeler is shown in Figure 3.

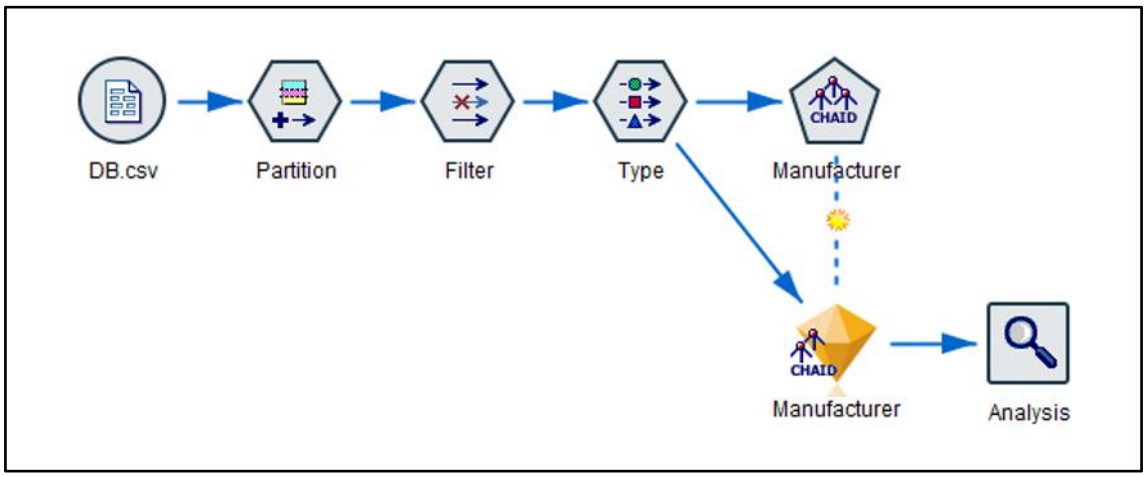

Figure 3 The overview of decision tree model processing in the SPSS Modeler.

2.5.4 Model Training and Evaluation

The training process was conducted using the decision tree node in SPSS Modeler. By linking the training data to the decision tree node, the model learned patterns within the data. After training, the following evaluation methods were used to assess the model's performance:

- Confusion Matrix: The confusion matrix was used to measure classification accuracy and analyze discrepancies between predicted and actual results.
- Accuracy: The overall accuracy of the model on the test data was calculated to ensure it had adequate predictive capability.
- ROC Curve: The ROC curve was plotted to further analyze the classification performance of the model, helping to select the optimal threshold.

3 Result

The decision tree analysis results from SPSS Modeler, shown in Figures 4 and 5, revealed the following key findings: To further understand the differences among manufacturers, the decision tree analysis provided key insights into the hierarchical structure of risk factors. The tree begins with the manufacturer as the root node, where different brands exhibit distinct risk profiles. The variables' importance determined the split criteria at each node, as indicated by the information gained in the C5.0 algorithm and chi-square statistics in the CHAID algorithm.

- Root Node – Manufacturer:

The first split was based on the vehicle manufacturer. Toyota, Honda, Nissan, Suzuki, and Mitsubishi vehicles significantly diverged in accident and defect trends. The model selected Toyota as the root node because it had the highest variance in risk factors compared to the other manufacturers.

- Second-Level Nodes – Mileage:

The next major split occurred for Toyota vehicles at mileage, specifically around the 80,000 km mark. Vehicles with higher mileage (above 80,000 km) were likelier to experience defects or accidents, particularly related to engine and transmission issues. In contrast, the split occurred at a lower mileage threshold for Suzuki and Nissan, indicating different usage patterns and durability levels.

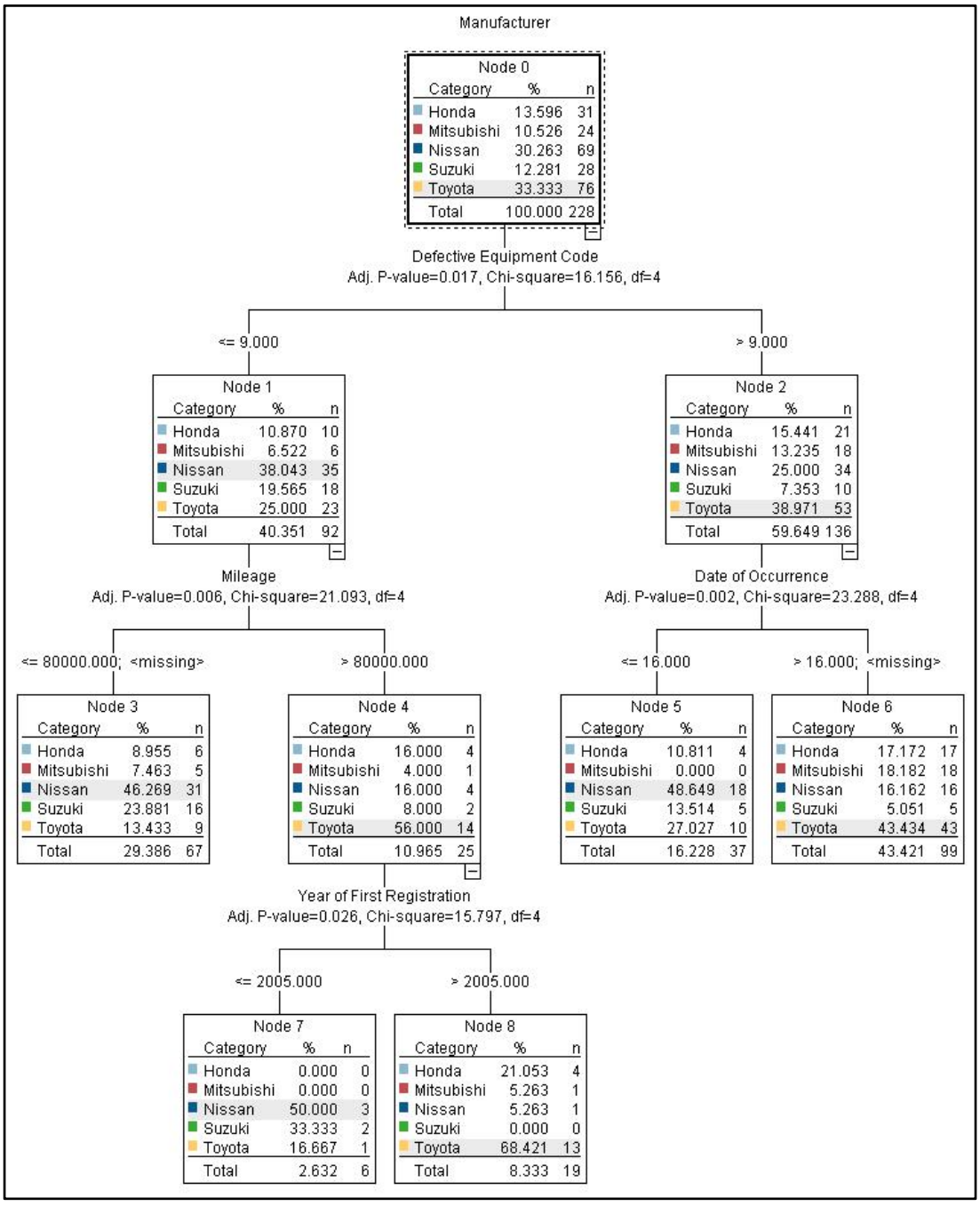

Figure 4 General view of decision tree analysis results with SPSS Modeler.

- Third-Level Nodes – First Registration Year:

The model further split Toyota vehicles based on the first registration year. Vehicles registered after 2006 exhibited fewer defects and accidents, likely due to improvements in manufacturing standards and safety regulations. However, the split based on the defective part code (codes below 9 or above 10) was more significant for Suzuki vehicles, revealing vulnerabilities in specific components.

- Leaf Nodes – Defective Part:

In the final splits, the decision tree revealed that vehicles with defective parts coded below 9 were primarily associated with Suzuki, while parts coded above 10 were more frequent in Toyota, Honda, and Mitsubishi. This division highlights the differences in component reliability across manufacturers, with some brands having recurring issues with specific parts.

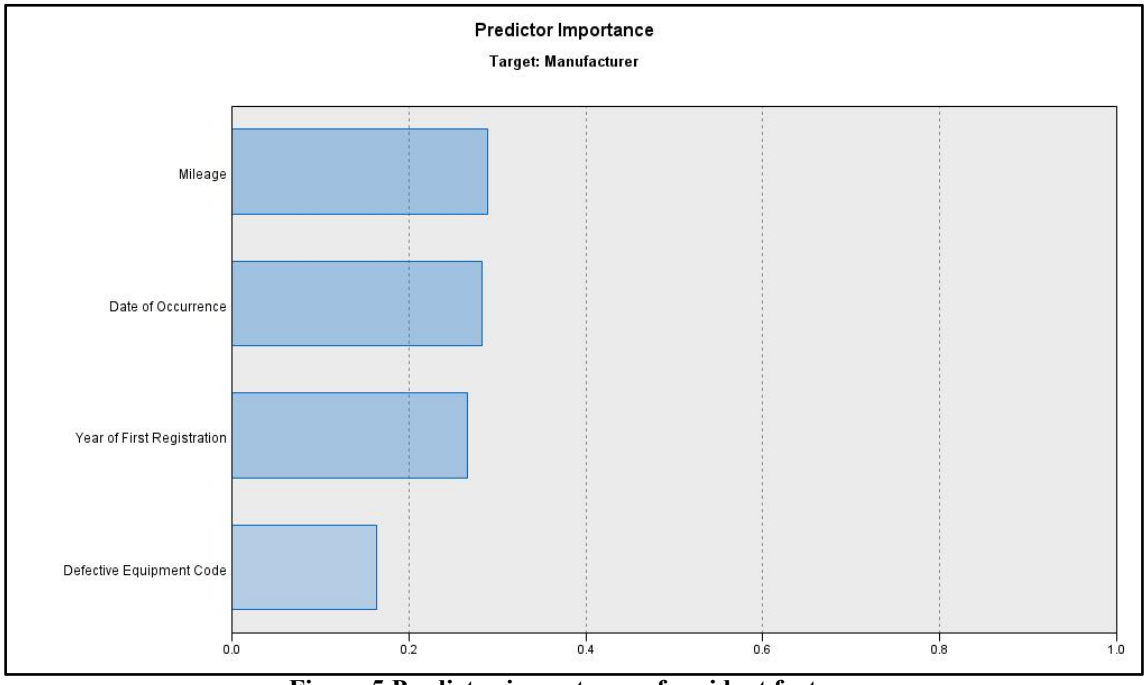

Figure 5 Predictor importance of accident factors.

3.1 Impact of Mileage

Main Observation: The occurrence of accidents or defects in vehicles showed variability across different manufacturers based on mileage.

Specific Results: Vehicles with over 80,000 km mileage were more likely to experience accidents or defects if they were Toyota. On the other hand, for vehicles with less than 80,000 km mileage, Suzuki and Nissan had a higher proportion of accidents or defects.

No Significant Change: In the case of Honda and Mitsubishi, there was no noticeable variation in accident or defect occurrence related to mileage.

The analysis revealed that vehicles with a mileage exceeding 80,000 km were more prone to accidents or defects in Toyota vehicles. Conversely, Suzuki and Nissan vehicles demonstrated a higher occurrence of accidents or defects at lower mileage, specifically below 80,000 km. The differences observed may be attributable to each manufacturer's distinct design, durability, and quality control processes, warranting further exploration in future studies.

3.2 Impact of First Registration Year

Main Observation: For Toyota vehicles, the frequency of accidents and defects varied based on the year of first registration.

Specific Results: Accidents and defects were more frequent in Toyota vehicles registered in or after 2006, with 13 cases for those registered before 2005 and only 1 case for those registered after 2006.

No Significant Change: Other manufacturers did not show a noticeable trend in relation to the year of first registration, likely due to the smaller number of cases.

For Toyota, vehicles registered after 2006 showed significantly fewer accidents and defects compared to those registered before 2005. This could suggest improvements in manufacturing processes or safety standards over time. However, no such trend was observed in other manufacturers, likely due to the smaller sample size for those cases.

3.3 Impact of Defective Equipment

Main Observation: The defective part significantly influenced the frequency of accidents or defects.

Specific Results: Overall, there were 92 cases with a defective part code below 9, and 136 cases with a code of 10 or higher. By manufacturer, Suzuki had more cases with defective part codes below 9, while Toyota, Honda, and Mitsubishi had more cases with codes of 10 or higher. Nissan did not show any significant changes based on the defective part code.

There were clear distinctions between manufacturers regarding the types of defective parts. Suzuki vehicles exhibited more defects in parts coded below 9, whereas Toyota, Honda, and Mitsubishi had a higher frequency of defects in parts coded 10 or above. These findings may point to manufacturer-specific weaknesses in particular vehicle components, which can inform targeted quality control and risk mitigation efforts.

These decision tree analysis results reveal different risk distributions across automotive manufacturers based on mileage, first registration year, and defective part, providing insights for targeted risk management.

4. Conclusion and Future Work

In this study, the integration of three types of automotive risk information databases and their analysis using decision trees were explored. The results indicated that the factors contributing to accidents and defects vary across different manufacturers. The decision tree analysis revealed important differences in accident and defect trends among manufacturers. For example, Toyota's vehicles with high mileage (over 80,000 km) were more susceptible to accidents or defects, whereas Suzuki and Nissan vehicles had more issues at lower mileage ranges. These patterns may be explained by differences in vehicle durability, usage patterns, or maintenance strategies across manufacturers. Understanding these distinctions is crucial for tailoring risk management strategies to specific manufacturers or vehicle types.

Additionally, the analysis identified the first registration year as a significant predictor for Toyota vehicles, with newer vehicles (post-2006) showing lower accident rates. This could be reflective of improvements in safety standards or vehicle technology. However, other manufacturers did not exhibit this trend, which could be due to variations in production methods or different regulatory environments.

The defective part codes provided further insights into manufacturer-specific vulnerabilities. Suzuki's high incidence of defects in parts coded below 9 indicates potential weaknesses in specific vehicle components, while Toyota and Honda had more defects in higher coded parts. These results could guide manufacturers in focusing on particular areas for quality improvement and safety enhancement.

Future research should incorporate larger datasets and additional manufacturers to ensure broader applicability of the findings. Moreover, expanding the analysis to include other machine learning techniques, such as Bayesian networks, could provide more nuanced insights into the causal relationships between vehicle risk factors.

Acknowledgement

I would like to express my sincere gratitude to Japan’s Ministry of Land, Infrastructure, Transport and Tourism (MLIT) for providing the automotive recall data used in this analysis.

References

[1] Japan Automobile Inspection and Registration Information Association Website, Statistical Information, "Trends in the Number of Vehicle Ownership (FY2022)."

[2] Ministry of Land, Infrastructure, Transport and Tourism Website, "Automobile Recall and Defect Information," "Trends in the Number of Automobile Recall Reports and Total Number of Affected Vehicles (FY2022)."

[3] Pourroostaei Ardakani, S., Liang, X., Mengistu, K. T., So, R. S., Wei, X., He, B., & Cheshmehzangi, A. (2023). Road car accident prediction using a machine-learning-enabled data analysis. Sustainability, 15(7), 5939.

[4] Atwah, A., & Al-Mousa, A. (2021, September). Car accident severity classification using machine learning. In 2021

International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT) (pp. 186-192). IEEE.

[5] Naoki Noguchi et al., "Exploring the Factors of Serious Traffic Accidents through Traffic Accident Data Analysis," Proceedings of the 37th Fuzzy System Symposium, pp. 675-680, 2021.

[6] AlKheder, S., AlRukaibi, F., & Aiash, A. (2023). Support vector machine (SVM), random forest (RF), artificial neural network (ANN) and Bayesian network for prediction and analysis of GCC traffic accidents. Journal of Ambient Intelligence and Humanized Computing, 14(6), 7331-7339.

[7] Aiash, A., & Robusté, F. (2023). Analyzing pedestrians' crash injury risk factors in Barcelona. Transportation research procedia, 71, 235-243.

[8] Song, Y. Y., & Ying, L. U. (2015). Decision tree methods: applications for classification and prediction. Shanghai archives of psychiatry, 27(2), 130.

[9] De Ville, B. (2013). Decision trees. Wiley Interdisciplinary Reviews: Computational Statistics, 5(6), 448-455.

[10] Gabriele Prati at al. Using data mining techniques to predict the severity of bicycle crashes. Accident Analysis and Prevention, 101 44-54, 2017.

[11] Han, Z., & Wen, L. (2022). Development and validation of a decision tree classification model for the essential hypertension based on serum protein biomarkers. Annals of Translational Medicine, 10(18).

[12] Luo, Z., Luo, B., Wang, P., Wu, J., Chen, C., Guo, Z., & Wang, Y. (2023). Predictive model of functional exercise compliance of patients with breast cancer based on decision tree. International Journal of Women's Health, 397-410.

[13] Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.

[14] Giri, A., & Paul, P. (2020). APPLIED MARKETING ANALYTICS USING SPSS: MODELER, STATISTICS AND AMOS GRAPHICS. PHI Learning Pvt. Ltd..

[15] Pandya, R., & Pandya, J. (2015). C5. 0 algorithm to improved decision tree with feature selection and reduced error pruning. International Journal of Computer Applications, 117(16), 18-21.

[16] Milanović, M., & Stamenković, M. (2016). CHAID decision tree: Methodological frame and application. Economic Themes, 54(4), 563-586.

[17] Liu, Y. Y., Yang, M., Ramsay, M., Li, X. S., & Coid, J. W. (2011). A comparison of logistic regression, classification and regression tree, and neural networks models in predicting violent re-offending. Journal of Quantitative Criminology, 27, 547-573.