

**Project Report**

**MACHINE LEARNING TECHNIQUES FOR PATTERN RECOGNITION APPLICATIONS**

**Artificial Intelligence (BEJ42803)**

**Dr. Nor Surayahani**

**Universiti Tun Hussein Onn Malaysia**

**08/07/2023**

**Semester 2 2023**

**Moritz Hoehnel (JD220004), Mattis Ritter (JD220003)**

Table of Contents

[1. Introduction 3](#_Toc138734430)

[2. Problem Statement 3](#_Toc138734431)

[3. Methodology 3](#_Toc138734432)

[4. Dataset Descriptions 3](#_Toc138734433)

[5. Results 3](#_Toc138734434)

[6. Analysis and Discussion 3](#_Toc138734435)

[7. Conclusion 3](#_Toc138734436)

[8. References 3](#_Toc138734437)

# Introduction

For transport companies the customer satisfaction is the most important issue. A satisfied customer is not just likelier to return, in times of social media word spreads fast about the quality of the transport [1]. This is why transport companies need to ensure that customers liked their traveling experience. This report focuses on analysing a customer survey. The survey considers different aspects of customer experience like the punctuality. To have more accurate results and consider the feelings of range of different customers, big data sets get analysed. As this is hardly possible with conventional methods pattern recognition will be used [2]. This report presents machine learning techniques to analyse the customer survey with pattern recognition.

The customer survey is about the Shinkansen Bullet Train which is Japan’s highspeed train. The objective of the analysis is to predict if the customer was satisfied with the journey. Therefor different prediction models shall be used. It is also important to understand the data and find out which parameter contributes at which weight to the customer satisfaction.

# Problem Statement

“In 2021, the estimated number of domestic travellers using Shinkansen high-speed railways in Japan totalled around 34.52 million” [3]. To find out if they had a satisfactory trip random people where ask to perform a survey. The results of this survey need to be analysed. A total of 94379 customer data with 13 categories each need to be looked at. The issue is to find a prediction to know if a customer is satisfied with the overall experience.

# Methodology

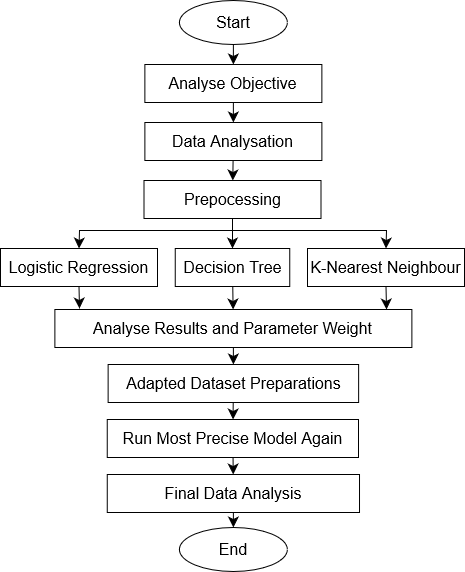


Figure 3.1: Methodology Flow Chart

The above figure gives an overview about the methodology of the project. The flow chart shows that after the project start the objective got analysed. For the following steps the project team used the Jupyter Notebook. After the dataset was loaded into the notebook, the set was analysed concerning its structure. This helped to prepare the dataset during preprocessing, where unsignificant parameters were removed, encoding was performed and training and test set were spilt apart. Afterwards the flow chart divides into three branches. This is where the first round of model running was performed. As far as the notebook is concerned, the models are fitted in the following order: Logistic Regressions, Decision Trees and K-Nearest Neighbour. As they rely on the same data set, they can be performed in any order. To find the best prediction for the data, the Logistic Regression was run with two different settings, while the Decision Tree was run with three and the K-Nearest Neighbour had one. In the step of ‘Analyse Results and Parameter Weight’ the first phase was scrutinise the output of the test set with focus on precisions. Based on the results the most precise model was chosen, which was one of the decision trees. The model was later used again. The second phase is to analyse which parameters have a high relative importance. Parameters with low importance got dropped in the following step of adapted data set preparation. Furthermore, two new approaches of data encoding were used, one by adding dummy values for all parameters another by encode null values with “-1”. As the decision tree model with criterion entropy and maximum depth 15 delivered the best test set results, it was run again. After that a final data analysis was performed. Here one can also find the comparison of the different model runs.

# Dataset Descriptions

The following chapter describes the dataset. To understand the dataset the functions “.head()”, “.tail()”, “.info()” and “.describe()”, had been applied as well as it was inspected for null values. The “.tail()” function reveals that the dataset has 94379 entires. The “.info()” function reveals that the dataset has 14 columns. The first one is the survey “ID” and the last one is the “Overall\_Experience” which evaluates if the passenger liked the journey. The customer can either like or dislike, which is expressed by a “1” or “0” respectively. To analyse the remaining categories figure 4.1 shows a summary of the most important dataset information.

|  |  |  |
| --- | --- | --- |
| # Column Dtype Null Count  ­--------------------------------------------------   |  |  | | --- | --- | | 0 ID int64  1 Departure\_Delay\_in\_Mins int64  2 Arrival\_Delay\_in\_Mins int64  3 Gender object  4 Seat\_Comfort object  5 Seat\_Class object  6 Arrival\_Time\_Convenient object  7 Onboard\_Wifi\_Service object  8 Ease\_of\_Online\_Booking object  9 Baggage\_Handling object  10 Legroom object  11 CheckIn\_Service object  12 Cleanliness object  13 Overall\_Experience int64 | 0  0  0  77  61  0  8930  30  73  142  90  77  6  0 | |

Figure 4.1: Summary of Dataset Description

Under “#” each column has a number aliened. The column “column” shows the names of the categories. “Dtype” is short for datatype. One can see that the datatype is either and integer or an object. For creating a model, all objects need to be changed to numbers, which is happening during preprocessing. “Null Count” is displaying the number of rows for each column that don not have a result inserted. Rows that have one or several empty fields need also treatment in the preprocessing.

# Results

Analyse der modelle und ihrer genauigkeit

# Analysis and Discussion

Analyse der confusion matrix

This chapter discusses the meaning of the model results. As Model 9, the decision tree with criterion entropy and maximum depth 15, scored the highest accuracy, it is chosen as the model of usage. The analysis will focus solely on model 9.

For transport companies it is important to know which categories are the most important for their customers. The Shinkansen team can focus to improve the experience for certain aspects. This then would lead to more customers being satisfied. To extract this information a Feature Importances bar graph was created.

Ein Bild, das Text, Screenshot, Diagramm, Schrift enthält.

Automatisch generierte Beschreibung

Figure 6.1 Feature Importances of Decision Tree

The bars displayed in figure 6.1 indicate the Relative Importance of each of the categories. The top five categories include “Seat\_Comfort”, “Ease\_of\_Online\_Booking”, “Arrival\_Time\_Convenient”, “Legroom” and “Baggage\_Handling”. It is also visible that the first two categories cumulated make up over 50% of the Relative Importance. This means that these two aspects need special attention by the transport company. It is also interesting to see that the two categories for delay have the lowest importance, despite Japanese are known for taking punctuality seriously. This can rule from the fact that at least 50% of the trains do not have a delay and if there is a delay it is only little. Only some trains have huge delays. The delay information have been acquired by using the “.describe()” function.

|  |
| --- |
|  |

Figure 6.2: Delay Information

As business recommendations the team would encourage the transport provider to keep special focus on their seats and their online booking system. This means for new trains to spend more money on comfortable seats that have as example a convenient head rest. But also service the seats regularly. As the online booking is also important to have an own tool that is easy to use, but also open the selling of tickets to other booking platforms, a customer may be already used to.

A further advice is to keep the punctuality at the level as it is know to not risk a decline in user experience.

# Conclusion

Wir sind sooooo geil

# References

|  |  |
| --- | --- |
| [1] | S. Kumar and M. Zymbler, “A machine learning approach to analyze customer satisfaction from airline tweets”, *springeropen.com*, Jul. 17, 2019. [Online]. Available: <https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0224-1> [Accessed: Jun. 27, 2023] |
| [2] | R. Szeliski, *Computer Vision*. Seattle: Springer Cham, 2022. |
| [3] | A. Arba, “Number of domestic Shinkansen travelers in Japan 2016-2021”, *statista.com*, Oct. 15, 2022. [Online]. Available: <https://www.statista.com/statistics/1272144/japan-tourism-domestic-shinkansen-traveler-number/> [Accessed: Jun. 27, 2023] |