An Automated Hybrid Approach for Feature Classification and Non-Numeric Feature Transformation in Medical AI domain

ZIHENG ZHANG[1], HAN FENG[1], MINGHUI TANG[2,3], KENJI HIRATA[2,3], GENKI HATA[4],
KATSUHIKO OGASAWARA[2,5], KOHSUKE KUDO[2,3], JUN NAKAYA[2,3]

[1]Graduate School of Medicine, Hokkaido University, Sapporo, Japan
[2]The Medical AI Research and Development Center, Hokkaido University, Sapporo, Japan
[3]Faculty of Medicine, Hokkaido University, Sapporo, Japan
[4]Products Development Department, Bathclin Corporation, Ibaraki, Japan
[5]Faculty of Health Sciences, Hokkaido University, Sapporo, Japan

Abstract

In current artificial intelligence (AI) researches utilizing structured data, machine learning algorithms often struggle to handle non-numeric features within structured data which limit their applicability. This challenge becomes particularly significant when dealing with medical or health-related data, as the structured data in medicine encompasses not only numerous numeric features but also various non-numeric features such as ordinal, categorical, datetime, and text features. Converting non-numeric features into a suitable format for machine learning algorithms is essential to facilitate model training. Conventionally, this process is done manually, which can be subjective, time-consuming and labor-intensive. To address this issue, the aim of our research is to proposes an innovative hybrid approach based on rules and large language models ChatGPT for automatic feature classification within structured data, and to convert non-numeric features into numeric representations that can be effectively utilized by machine learning algorithms. The automation aspect of our method holds great promise in ensuring efficient and accurate handling of diverse non-numeric data. By leveraging the power of large language models, we expect our approach to streamline the preprocessing of medical and health-related data and enhance the performance of AI models in clinical applications.

Introduction

In machine learning, a feature is an individual measurable property or characteristic of a phenomenon. Choosing informative, discriminating and independent features is a crucial element of effective algorithms in pattern recognition, classification and regression. When undertaking machine learning tasks, meticulous feature selection and effective feature engineering are fundamental determinants for achieving desirable results and augmenting model performance.

However, in current artificial intelligence (AI) researches utilizing structured data, machine learning algorithms often struggle to handle non-numeric features within structured data which limit their applicability. This challenge becomes particularly significant when dealing with medical or health-related data, as the structured data in medicine encompasses not only numerous numeric features but also various non-numeric features such as ordinal, categorical, datetime, and text features. Converting non-numeric features into a suitable format for machine learning algorithms is essential to facilitate model training, as these non-numeric features in medical data are likely to contain critical intelligence that can affect outcomes. Conventionally, this process is done manually, which can be subjective, time-consuming and labor-intensive.

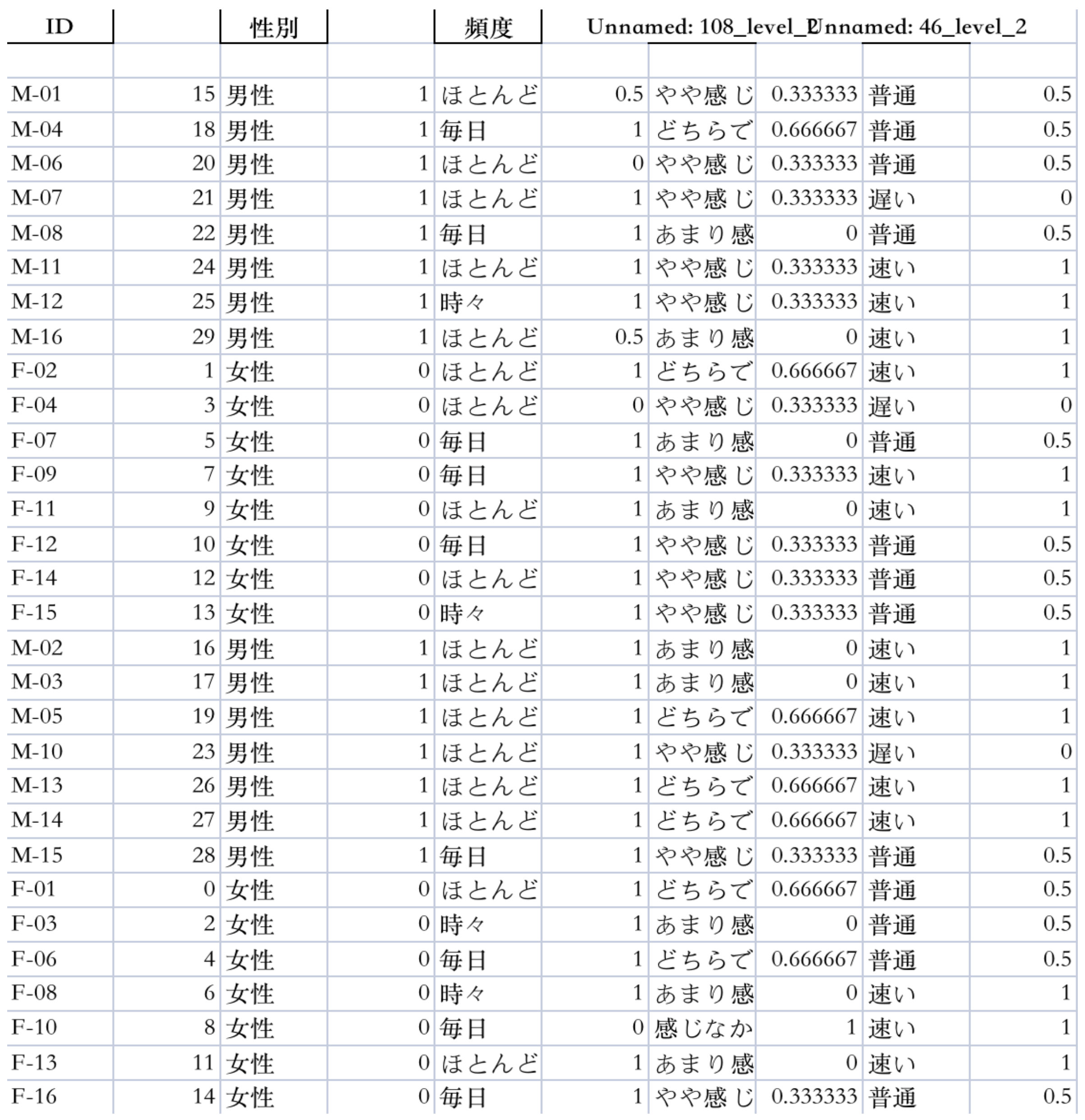
Objective

In this study, we aim to proposes an innovative hybrid approach based on rules and large language models GPT for automatic feature classification within structured data, and to convert non-numeric features into numeric features that can be effectively utilized by machine learning algorithms.

Results

- First, based on the rule-based approach, we successfully categorized the 258 features of 30 objects into 192 numeric features and 66 non-numeric features.
- Next, utilizing the GPT model, we further classified these 66 non-numeric features into ordinal and categorical features, resulting in 20 ordinal features and 46 categorical features. The categorical features were assigned values using Label Encoding. For the ordinal features, we utilized GPT once more to reorder them, mapping them to the [0, 1] interval based on linear relationships.

Figure 3: Example of converting non-numeric features into numeric features



ID		性別		頻度		Unnamed: 108_level_2		Unnamed: 46_level_2	
M-01	15	男性	1	ほとんど	0.5	やや感じ	0.333333	普通	0.5
M-04	18	男性	1	毎日	1	どちらで	0.666667	普通	0.5
M-06	20	男性	1	ほとんど	0	やや感じ	0.333333	普通	0.5
M-07	21	男性	1	ほとんど	1	やや感じ	0.333333	遅い	0
M-08	22	男性	1	毎日	1	あまり感	0	普通	0.5
M-11	24	男性	1	ほとんど	1	やや感じ	0.333333	速い	1
M-12	25	男性	1	時々	1	やや感じ	0.333333	速い	1
M-16	29	男性	1	ほとんど	0.5	あまり感	0	速い	1
F-02	1	女性	0	ほとんど	1	どちらで	0.666667	速い	1
F-04	3	女性	0	ほとんど	0	やや感じ	0.333333	遅い	0
F-07	5	女性	0	毎日	1	あまり感	0	普通	0.5
F-09	7	女性	0	毎日	1	やや感じ	0.333333	速い	1
F-11	9	女性	0	ほとんど	1	あまり感	0	速い	1
F-12	10	女性	0	毎日	1	やや感じ	0.333333	普通	0.5
F-14	12	女性	0	ほとんど	1	やや感じ	0.333333	普通	0.5
F-15	13	女性	0	時々	1	やや感じ	0.333333	普通	0.5
M-02	16	男性	1	ほとんど	1	あまり感	0	速い	1
M-03	17	男性	1	ほとんど	1	あまり感	0	速い	1
M-05	19	男性	1	ほとんど	1	どちらで	0.666667	速い	1
M-10	23	男性	1	ほとんど	1	やや感じ	0.333333	遅い	0
M-13	26	男性	1	ほとんど	1	どちらで	0.666667	速い	1
M-14	27	男性	1	ほとんど	1	どちらで	0.666667	速い	1
M-15	28	男性	1	毎日	1	やや感じ	0.333333	普通	0.5
F-01	0	女性	0	ほとんど	1	どちらで	0.666667	普通	0.5
F-03	2	女性	0	時々	1	あまり感	0	普通	0.5
F-06	4	女性	0	毎日	1	どちらで	0.666667	普通	0.5
F-08	6	女性	0	時々	1	あまり感	0	速い	1
F-10	8	女性	0	毎日	0	感じなか	1	速い	1
F-13	11	女性	0	ほとんど	1	あまり感	0	速い	1
F-16	14	女性	0	毎日	1	やや感じ	0.333333	普通	0.5

Data and Settings

- Data Source: Bathclin Corporation
- Sample Size: 30
- Features Size: 258
- GPT Version: gpt-3.5-turbo
- Temperature: 0

Methods

Figure 1: Classification of Numeric and Non-numeric Features Based on Rules

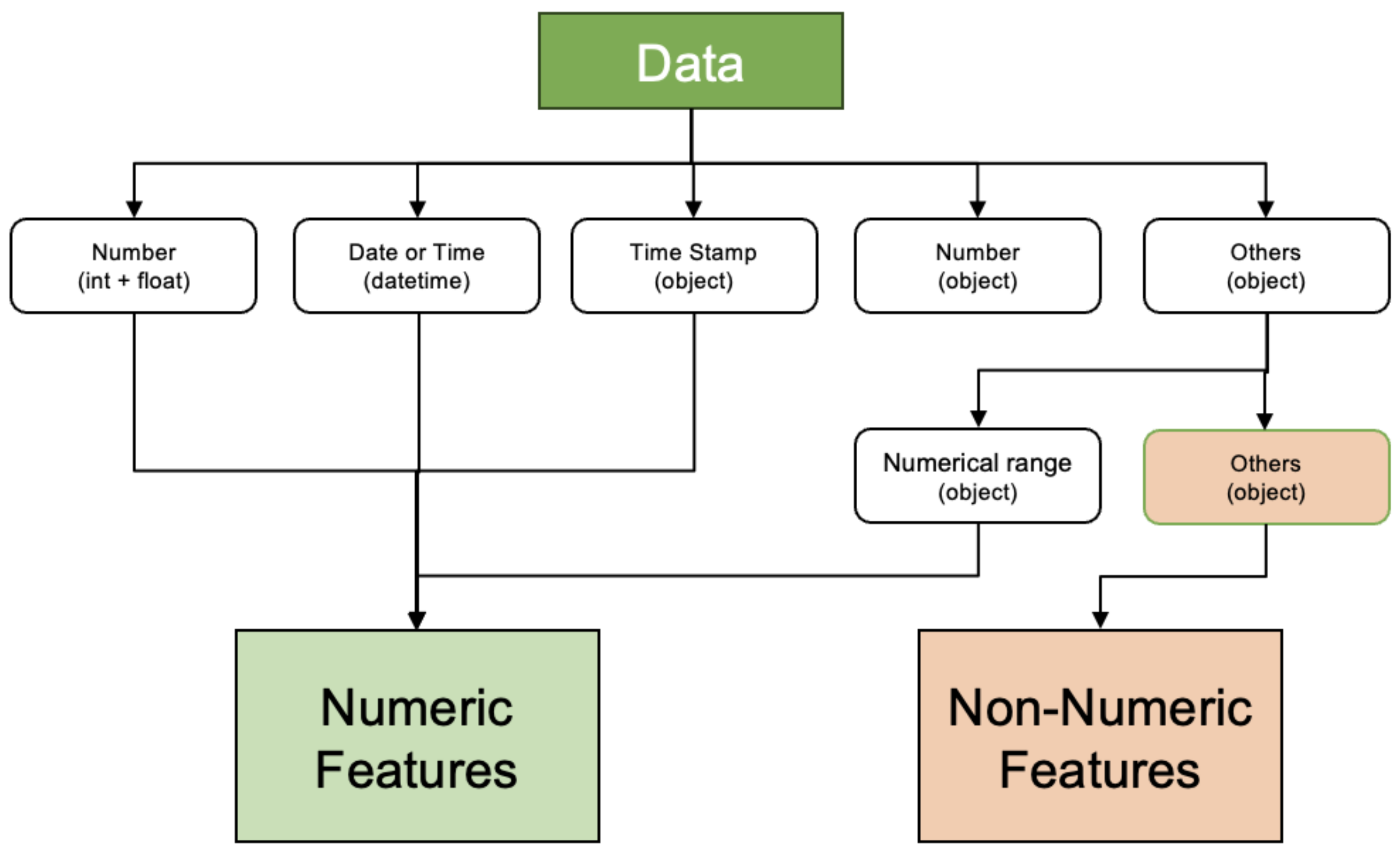
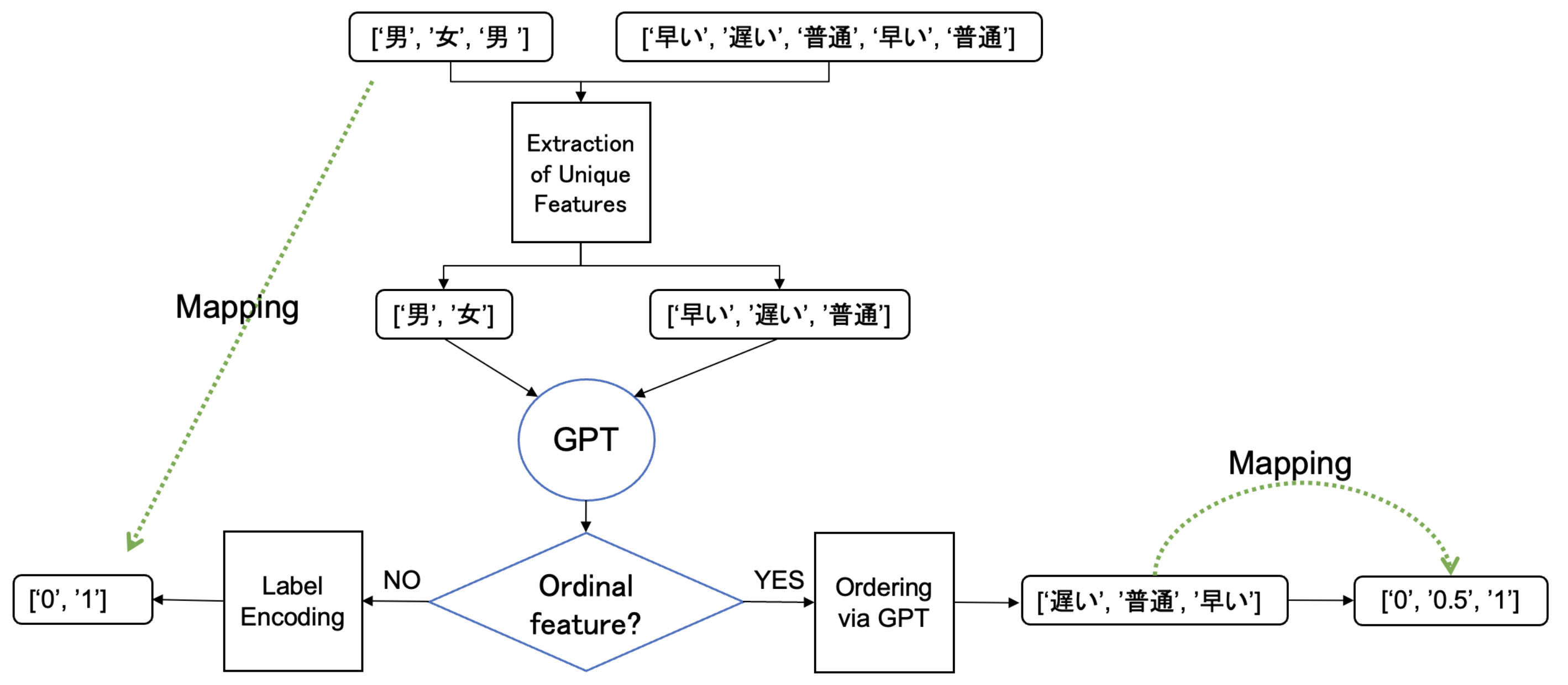


Figure 2: Example of Automated Classification Using GPT



Discussion & Conclusion

- The automated numerical conversion method employed in this study ensures the efficient and accurate processing of a wide array of non-numeric features while simultaneously avoiding human errors and subjectivity.
- Leveraging the power of large language models streamlines the preprocessing of medical and health-related data, holding immense promise to enhance the performance of AI models across various clinical and health applications, marking a significant stride towards improved healthcare outcomes and advancements in artificial intelligence

Limitations

- Up to this point, this methods can only handle structured data.
- The numerical features of ordinal features involved assigning values in the range [0, 1] based on a linear order. However, this linear relationship may not be applicable in cases where such an order is not present.