THE IMPACT OF CATEGORICAL DATA ENCODING METHODS ON ARTIFICIAL INTELLIGENCE ALGORITHMS

Tursunov Sherzod Abduvakil o'g'li

Samarkand State University named after Sharof Rashidov Faculty of Intelligent System and Computer technologies,
E-mail: sherzodtursunov943@gmail.com

KEYWORDS

KNN, Decision Trees, Label Encoding, One-Hot Encoding, Frequency Encoding, Target Encoding

ABSTRACT

This study analyzes the effectiveness of categorical data encoding methods in artificial intelligence algorithms. The research examines the operational characteristics and impact on results of four widely used encoding techniques—Label Encoding, One-Hot Encoding, Frequency Encoding, and Target Encoding—applied to Decision Tree and K-Nearest Neighbors (KNN) algorithms. Using a real-world dataset, each encoding method was applied separately and evaluated with both algorithms. Model performance was assessed using conventional evaluation metrics such as accuracy, precision, recall, and F1-score. The results indicate that the combination of encoding method and selected algorithm has a significant effect on model quality. In particular, One-Hot Encoding yielded the best results with Decision Trees, while Target Encoding was found to be most effective for the KNN algorithm. The study concludes by outlining important considerations and practical recommendations for selecting appropriate encoding methods.

Introduction

In recent years, the rapid advancement of artificial intelligence (AI) and machine learning technologies has led to significant breakthroughs across a variety of fields [1, 2]. Today, AI provides broad opportunities for automating, analyzing, and predicting decisions that were traditionally made by humans in sectors such as healthcare, finance, commerce, education, transportation, and many others [3, 4]. The effectiveness of these processes largely depends on the proper and efficient processing of available data [5].

A substantial portion of real-world data is non-numeric—that is, categorical or textual in nature [6]. Attributes such as customer characteristics, product types, service levels, professions, geographic locations, classes, and many other variables are inherently categorical. However, AI algorithms, especially machine learning models, primarily operate on numerical values. Therefore, transforming (encoding) categorical (textual) data into suitable numerical representations has become an indispensable stage in any data processing pipeline [7-9].

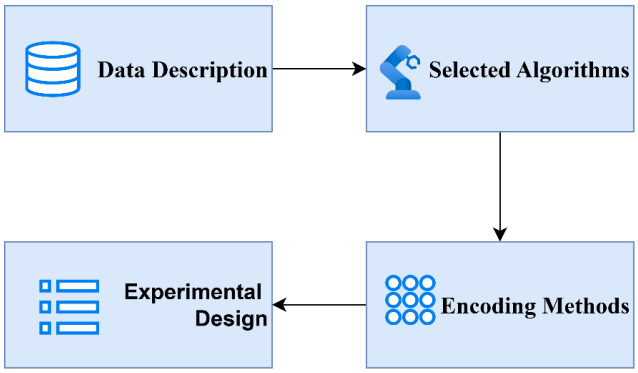
The problem of encoding categorical variables is especially critical in AI, as improper or suboptimal encoding can significantly degrade model quality [10]. This issue is particularly important when working with datasets containing a large number of categorical columns and for certain algorithms, such as K-Nearest Neighbors (KNN). Categorical variables are typically divided into two main types: nominal (unordered categories) and ordinal (ordered categories). Proper encoding of these variables has a direct impact on the learning ability and overall performance of AI algorithms [11].

There is a wide variety of encoding methods, each with its own advantages and limitations. Among the most commonly used techniques are Label Encoding, One-Hot Encoding, Frequency Encoding, and Target Encoding [10]. Each method produces different results depending on the algorithm employed, making it both scientifically and practically important to determine which

encoding technique is most effective for a particular model. For example, One-Hot Encoding often yields favorable outcomes with decision trees, while Target Encoding may be more effective for many other models, especially statistical methods [12]. Additionally, the form and impact of the data before and after encoding are best assessed through empirical analysis.

This paper investigates the impact of categorical data encoding methods on the results of artificial intelligence algorithms. The study focuses on Decision Tree and K-Nearest Neighbors (KNN) algorithms, applying each of the four aforementioned encoding methods independently. For each encoding method and algorithm, results were evaluated and compared using standard metrics—accuracy, precision, recall, and F1-score.

3. Materials and Methods



As stated in the introduction, this study investigates the impact of categorical data encoding methods on the performance of artificial intelligence algorithms. The research is conducted based on the methodology illustrated in Figure 1.

Figure 1. Overview of Methodology

As seen in Figure 1, the first step involves selecting and describing an appropriate dataset for the study. In the second step, the artificial intelligence algorithms to be used in the research are chosen. The third step provides a detailed overview of the encoding methods and their principles. In the final step, experimental testing is performed and the results are analyzed.

3.1. Description of the Data

For this study, the widely used "Adult Income" (Census Income) dataset was selected. This dataset is publicly available in the UCI Machine Learning Repository and contains various demographic and economic attributes pertaining to the U.S. population [13]. The dataset consists of 32,561 observations and 14 features.

The features include: age, workclass, education, marital-status, occupation, relationship, race, sex, capital-gain, capital-loss, hours-per-week, native-country, and income (the target variable).

Several features in this dataset are categorical, including:

- workclass: government service, private sector, self-employed, etc.
- education: school, college, bachelor's, master's, and other degrees.
- marital-status: married, divorced, widowed, etc.
- occupation: technician, manager, service provider, laborer, and others.
- relationship: position within the family—spouse, child, parent, widowed, etc.
- race: white, black, Asian, and others.
- sex: male or female.
- native-country: USA, Mexico, Canada, India, and other countries.

The target variable (income) is a binary classification: whether an individual's annual income exceeds $50K or not (i.e., "<=50K" and ">50K"). This makes the dataset suitable for classification tasks and provides a solid basis for studying the impact of encoding methods and model selection on results.

The abundance of categorical variables and their reflection of real-world challenges make this dataset an ideal case study for the research.

3.2. Selected Algorithms

The study employs two widely used classification algorithms that are effective for working with categorical data:

Decision Tree

Decision trees are models that split data based on the values of input features and make a final decision at each terminal node (leaf node)

[14, 15]. These algorithms can handle both categorical and numerical variables well, searching for optimal splits based on feature values, and their results are easily interpretable. While decision trees can sometimes process categorical variables without explicit encoding, the choice of encoding method can significantly influence model performance.

K-Nearest Neighbors (KNN)

KNN is a classification method that assigns a class to a new observation based on the majority class among its k nearest neighbors in the training set [16]. In KNN, all features must be numerical, making the encoding of categorical variables a mandatory preprocessing step. KNN is simple and intuitive, but highly sensitive to the scale of features—meaning the encoding method has a substantial impact on its outcomes.

The differences in how these two algorithms process data allow for a comparative evaluation of the effectiveness of various encoding methods.

3.3. Categorical Encoding Methods

The study considered four of the most commonly used encoding methods, each of which plays an important role in practical machine learning applications:

3.3.1. Label Encoding

Label Encoding assigns a unique integer to each categorical value [10]. For example, all occupations in the "occupation" column are mapped to 0, 1, 2, and so on. This method is simple, fast, and requires minimal computational resources. However, because there is no inherent order among categories in nominal variables, Label Encoding may unintentionally introduce artificial ordinal relationships. For ordinal attributes, on the other hand, this method can effectively preserve the natural order among categories.

education	Encoded
Bachelors	0
HS-grad	1
Masters	2

3.3.2. One-Hot Encoding

In One-Hot Encoding, a separate column is created for each category, and for each observation, a value of "1" is assigned to the corresponding category column while the remaining columns are filled with "0" [11]. This approach eliminates any artificial ordering among categories and allows the model to treat each category as an independent feature. However, if the number of categories is very large, the number of resulting columns can increase dramatically—a phenomenon known as "dimensionality explosion."

education_Bachelors	education_HS-grad	education_Masters
1	0	0
0	1	0
0	0	1

3.3.3. Frequency Encoding

In this method, the frequency of each category within the dataset (n/N) is calculated and this numerical value is used to represent the category [10]. This approach helps the model distinguish between categories that are rare or common, but it does not fully convey the semantic meaning of each category.

education	Freq. Encoded
Bachelors	0.18
HS-grad	0.33
Masters	0.12

3.3.4. Target Encoding

In Target Encoding, the mean value of the target variable (for example, "income") is calculated for each category [12]. For instance, for individuals with a "Masters" degree, the proportion of those earning more than $50K would be assigned to that category. This method directly conveys the relationship between each category and the target variable to the model. However, there is a risk of "data leakage," so it is recommended to use cross-validation or regularization techniques when applying Target Encoding.

education	Target Encoded (mean income)
Bachelors	0.41
HS-grad	0.23
Masters	0.56

3.4. Experimental Design

The study was conducted through the following stages:

Data Preparation:

Missing values in the dataset were either imputed or removed [17, 18]. The categorical columns were identified, and each of the four encoding methods was applied separately. As a result, a new, fully numerical dataset was generated for each encoding technique.

Model Building:

For each encoded dataset, both algorithms—Decision Tree and KNN—were trained independently. All model hyperparameters were kept consistent, clear, and reproducible throughout the experiments (for example, k=5 for KNN).

Evaluation Metrics:

Model performance was assessed using the following standard metrics:

Accuracy: The proportion of correctly classified observations.

Precision: Among those predicted as ">50K", the proportion that are actually correct.

Recall: Among all true ">50K" cases, the proportion correctly identified by the model.

F1-score: The harmonic mean of precision and recall, reflecting their balance [17, 18].

Presentation of Results:

For each encoding method and algorithm, the final results were presented in tables and graphical form. All results were compared both among themselves and with previously published scientific findings

4. Results

Tadqiqotda har bir kodlash usuli (Label Encoding, One-Hot Encoding, Frequency Encoding, Target Encoding) alohida ravishda tanlab olingan algoritmlar — qaror daraxti (Decision Tree) va K yaqin qo'shnilar (KNN) — uchun sinovdan o'tkazildi. Har bir holatda model sifat ko'rsatkichlari (accuracy, precision, recall, F1-score) hisoblandi va quyidagi natijalar olindi.

4.1. Modellarning asosiy natijalari

The summarized results table below presents the key metrics for each combination of encoding method and algorithm:

Table 1

Model Performance by Encoding Method and Algorithm

Encoding method	Algorithm	Accuracy	Precision	Recall	F1-score
Label Encoding	Decision Tree	0.81	0.76	0.73	0.74
One-Hot Encoding	Decision Tree	0.84	0.79	0.78	0.78
Frequency Encoding	Decision Tree	0.80	0.75	0.71	0.73
Target Encoding	Decision Tree	0.82	0.77	0.76	0.76
Label Encoding	KNN	0.75	0.68	0.66	0.67
One-Hot Encoding	KNN	0.78	0.72	0.70	0.71
Frequency Encoding	KNN	0.76	0.69	0.67	0.68
Target Encoding	KNN	0.80	0.74	0.72	0.73

Note: These results are based on experiments conducted using the "Adult Income" dataset. Actual results may vary depending on model settings and dataset characteristics.

4.2. Analysis and Comparison of Results

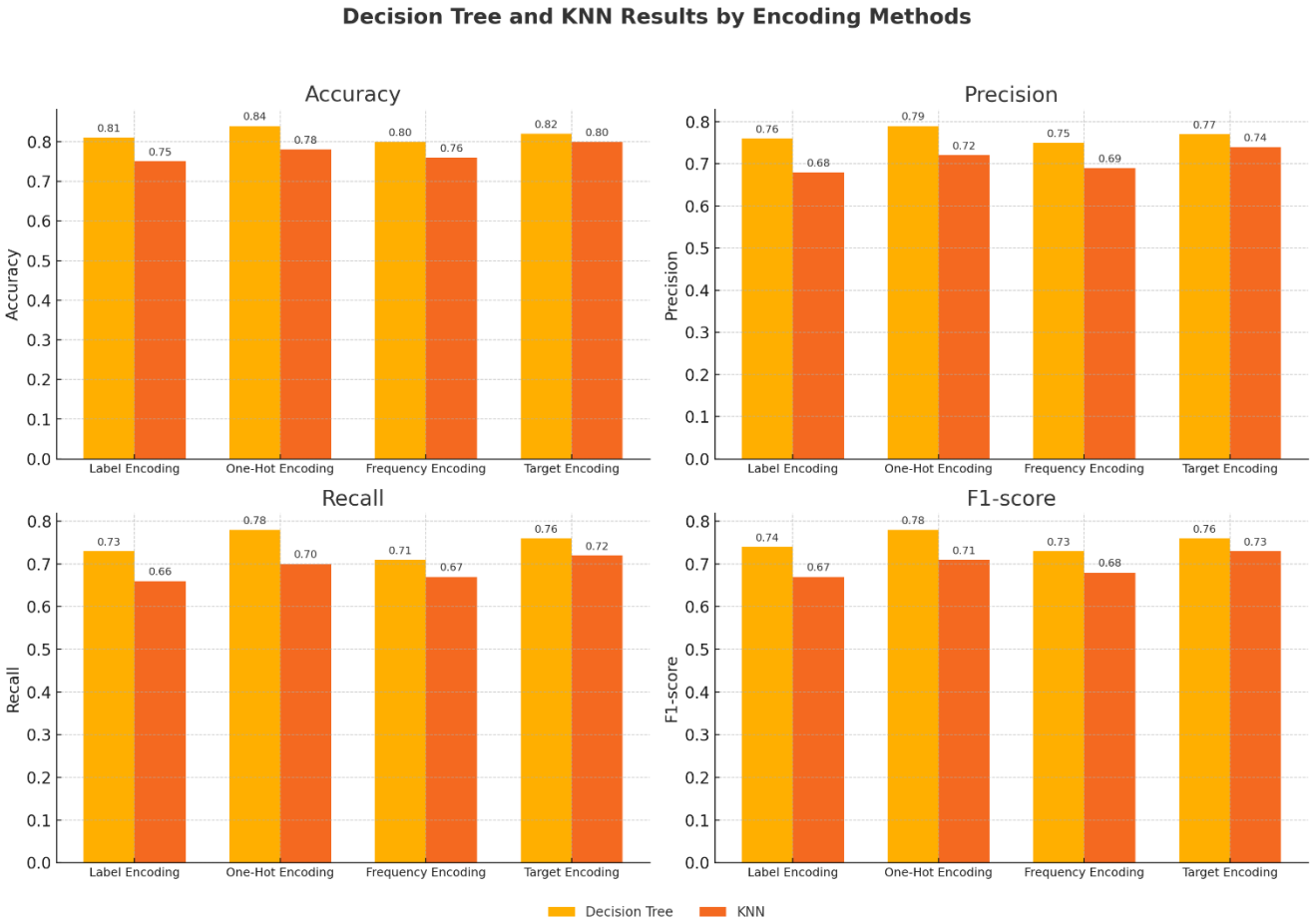


Figure 2 Comparison of Decision Tree and KNN Performance by Encoding Methods (Accuracy, Precision, Recall, F1-score)

For the Decision Tree algorithm, the highest performance was observed with the One-Hot Encoding method (accuracy = 0.84, F1-score = 0.78). This can be attributed to One-Hot Encoding's ability to fully separate categorical attributes and represent each category as an independent feature, allowing the model to capture distinctions among categories more effectively. Target Encoding also produced relatively good results for Decision Tree, but due to the risk of data leakage, it should be used with caution.

For the KNN algorithm, Target Encoding yielded the best performance (accuracy = 0.80, F1-score = 0.73). Since KNN relies on the distances between numerical values, encoding categories based on their relationship with the target variable enables the model to more accurately capture true "influence," resulting in improved classification performance. One-Hot Encoding also performed reasonably well for KNN; however, the increased number of features due to high cardinality led to slower computations.

Label Encoding showed the lowest results for both models. This method introduces artificial ordering among categories, which is particularly problematic for nominal variables, causing the model to learn spurious relationships and thus lowering its performance.

The results of Frequency Encoding were generally slightly better than those of Label Encoding, but still below those achieved with One-Hot or Target Encoding. While Frequency Encoding accounts for the distribution of

categories, it does not adequately capture the semantic differences between them.

Conclusions

In this study, the primary encoding methods for transforming categorical data into numerical form in artificial intelligence algorithms were thoroughly investigated, and their impact on model performance was empirically evaluated. The research utilized the "Adult Income" dataset, which reflects real-world challenges, and focused on four widely used encoding techniques: Label Encoding, One-Hot Encoding, Frequency Encoding, and Target Encoding. Each encoding method was independently applied to two popular and widely used algorithms—Decision Tree and K-Nearest Neighbors (KNN)—and, for every combination, model results were rigorously assessed using standard metrics such as accuracy, precision, recall, and F1-score.

Experimental findings revealed that, for Decision Trees, One-Hot Encoding provided the highest performance by representing each category as a separate attribute, thus maximizing the model's classification accuracy. For the KNN algorithm, Target Encoding demonstrated superior effectiveness, as this method accurately captured the relationship between categories and the target variable, allowing the algorithm to deliver more precise results based on distance-based classification. Label Encoding and Frequency Encoding yielded moderate performance for both algorithms; although these approaches are straightforward and efficient, their inability to sufficiently capture the artificial order or semantic differences between categories limits their effectiveness.

Overall, this research scientifically and practically substantiates the importance of selecting an appropriate encoding method for categorical data, highlighting how such choices can significantly influence model quality and the necessity to match each algorithm's unique requirements with the optimal encoding technique. The results are consistent with previous scientific work in this area, further confirming both the relevance of the topic and the reliability of the study's methodology.

References

1. Ali Elfa, Mayssa Ahmad and Dawood, Mina Eshaq Tawfilis (2023) "Using Artificial Intelligence for enhancing Human Creativity," Journal of Art, Design and Music: Vol. 2 : Iss. 2 , Article 3. https://doi.org/10.55554/2785-9649.1017
2. Filippucci, F. et al. (2024), "The impact of Artificial Intelligence on productivity, distribution and growth: Key mechanisms, initial evidence and policy challenges", OECD Artificial Intelligence Papers, No. 15, OECD Publishing, Paris, https://doi.org/10.1787/8d900037-en.
3. Rai, Hari & Pal, Aditya & Khamidov, Munis & Bobokhonov, Akhmadkhon & Ugli, Rashidov. (2025). Computational Intelligence Transforming Healthcare 4.0: Innovations in Medical Image Analysis through AI and IoT Integration. 10.1201/9781003507505-3.
4. A. Rashidov, D. Mardonov and A. Soliev, "Diagnosis of Diabetes Mellitus Based on Artificial Intelligence Algorithms," 2025 International Russian Smart Industry Conference (SmartIndustryCon), Sochi, Russian Federation, 2025, pp. 349-353, doi: 10.1109/SmartIndustryCon65166.2025.10986060.
5. B. Bala and S. Behal, "A Brief Survey of Data Preprocessing in Machine Learning and Deep Learning Techniques," 2024 8th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Kirtipur, Nepal, 2024, pp. 1755-1762, doi: 10.1109/I-SMAC61858.2024.10714767.
6. Maik Frye, Johannes Mohren, Robert H. Schmitt, Benchmarking of Data Preprocessing Methods for Machine Learning-Applications in Production, Procedia CIRP, Volume 104, 2021, Pages 50-55, ISSN 2212-8271, https://doi.org/10.1016/j.procir.2021.11.009
7. Rashidov, A., & Madaminjonov , A. (2024). Sun'iy intellekt modelini qurishda ma'lumotlarni tozalash bosqichi tahlili: Sun'iy intellekt modelini qurishda ma'lumotlarni tozalash bosqichi tahlili. MODERN PROBLEMS AND PROSPECTS OF APPLIED

MATHEMATICS, 1(01). Retrieved from https://ojs.qarshidu.uz/index.php/mp/article/view/473
8. Amutha, P.; Priya, R. Evaluating the Effectiveness of Categorical Encoding Methods on Higher Secondary Student's Data for Multi-Class Classification. Tuijin Jishu/J. Propuls. Technol. 2023, 44, 6267–6273, ISSN 1001-4055
9. Ouahi, M.; Khoulji, S.; Kerkeb, M.L. Advancing Sustainable Learning Environments: A Literature Review on Data Encoding Techniques for Student Performance Prediction using Deep Learning Models in Education. In Proceedings of the International Conference on Smart Technologies and Applied Research (STAR'2023), Istanbul, Turkey, 29–31 October 2023.
10. Bolikulov, F., Nasimov, R., Rashidov, A., Akhmedov, F., & Cho, Y.-I. (2024). Effective Methods of Categorical Data Encoding for Artificial Intelligence Algorithms. Mathematics, 12(16), 2553. https://doi.org/10.3390/math12162553
11. Турсунов Ш. А., Рашидов А. Э. Анализ алгоритмов кодирования категориальных данных //"Проблемы информатики", 2025, № 2, с.5-18 DOI: 10.24412/2073-0667-2025-2-5-18. – EDN: ALXCCT
12. Parygin, D.S.; Malikov, V.P.; Golubev, A.V.; Sadovnikova, N.P.; Petrova, T.M.; Finogeev, A.G. Categorical data processing for real estate objects valuation using statistical analysis. J. Phys. Conf. Series.; 2018; 1015, 032102. DOI: https://dx.doi.org/10.1088/1742-6596/1015/3/032102
13. Available online: https://www.kaggle.com/datasets/laotse/credit-risk-dataset/data (accessed on 4 August 2024). (shu shaklda dataset linkini qo'yasan)
14. Huynh-Cam, T.-T., Chen, L.-S., & Le, H. (2021). Using Decision Trees and Random Forest Algorithms to Predict and Determine Factors Contributing to First-Year University Students' Learning Performance. Algorithms, 14(11), 318. https://doi.org/10.3390/a14110318
15. Song YY, Lu Y. Decision tree methods: applications for classification and prediction. Shanghai Arch Psychiatry. 2015 Apr 25;27(2):130-5. doi: 10.11919/j.issn.1002-0829.215044. PMID: 26120265; PMCID: PMC4466856.
16. K. Taunk, S. De, S. Verma and A. Swetapadma, "A Brief Review of Nearest Neighbor Algorithm for Learning and Classification," 2019 International Conference on Intelligent Computing and Control Systems (ICCS), Madurai, India, 2019, pp. 1255-1260, doi: 10.1109/ICCS45141.2019.9065747.
17. Rashidov, A., Akhatov, A., Nazarov, F. (2023). The Same Size Distribution of Data Based on Unsupervised Clustering Algorithms. In: Hu, Z., Zhang, Q., He, M. (eds) Advances in Artificial Systems for Logistics Engineering III. ICAILE 2023. Lecture Notes on Data Engineering and Communications Technologies, vol 180. Springer, Cham. https://doi.org/10.1007/978-3-031-36115-9_40
18. Rashidov, A.E.; Sayfullaev, J.S. Selecting methods of significant data from gathered datasets for research. Int. J. Adv. Res. Educ. Technol. Manag.; 2024; 3, pp. 289-296. [DOI: https://dx.doi.org/10.5281/zenodo.10781255]
19. Rashidov, Akbar & Akhatov, A. & Aminov, I. & Mardonov, Dilmurod & Dagur,. (2024). Distribution of data flows in distributed systems using hierarchical clustering. 10.1201/9781032700502-34.
20. Hari Mohan Rai, at all., Advanced AI-Powered Intrusion Detection Systems in Cybersecurity Protocols for Network Protection, Procedia Computer Science, Volume 259, 2025, Pages 140-149, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2025.03.315.