Effective Methods of Categorical Data Encoding for Artificial Intelligence Algorithms

Furkat Bolikulov 1, Rashid Nasimov 2, Akbar Rashidov 3, Farkhod Akhmedov 1,*, and Young-Im Cho 1,*,*

1 Department of Computer Engineering, Gachon University, Sujeong-gu, Seongnam-si 461-701, Gyeonggi-do, Republic of Korea; bolikulovfurkat@mail.ru
2 Department of Information Systems and Technologies, Tashkent State University of Economics, Tashkent 100066, Uzbekistan; rashid.nasimov@tsue.uz
3 Department of Artificial Intelligence and Information Systems, Samarkand State University Named after Sharof Rashidov, Samarkand 140100, Uzbekistan; researcher.are@samdu.uz
* Correspondence: farhod34@gachon.ac.kr (F.A.); yicho@gachon.ac.kr (Y.-I.C.)

Abstract: It is known that artificial intelligence algorithms are based on calculations performed using various mathematical operations. In order for these calculation processes to be carried out correctly, some types of data cannot be fed directly into the algorithms. In other words, numerical data should be input to these algorithms, but not all data in datasets collected for artificial intelligence algorithms are always numerical. These data may not be quantitative but may be important for the study under consideration. That is, these data cannot be thrown away. In such a case, it is necessary to transfer categorical data to numeric type. In this research work, 14 encoding methods of transforming of categorical data were considered. At the same time, conclusions are given about the general conditions of using these methods. During the research, categorical data in the dataset that were collected in order to assess whether it is possible to give credit to customers will be transformed based on 14 methods. After applying each encoding method, experimental tests are conducted based on the classification algorithm, and they are evaluated. At the end of the study, the results of the experimental tests are discussed and research conclusions are presented.

Keywords: artificial intelligence; data preprocessing; encoding methods; logistic classification algorithm; classification assessment methods

MSC: 68T01

1. Introduction

Today, scientists of the world are conducting many research projects dedicated to finding solutions to various problems with the help of artificial intelligence algorithms and models. As a result of these ongoing and completed studies, new artificial intelligence algorithms are being developed or existing ones are being improved [1–5]. As a result, the indicators of accuracy in finding a solution to a given problem are improving [6,7]. However, the accuracy in this field depends not only on the mathematical models, the algorithms of artificial intelligence based on them, but also on the quality indicator of the data included in the algorithm of artificial intelligence [8,9]. In other words, correctly selected and preprocessed data are a prerequisite for efficiency.

The data preprocessing is carried out in several stages depending on the collected data, the problem, and the applied artificial intelligence algorithms [10]. However, the most commonly performed preliminary data preprocessing steps before building artificial intelligence models are the following [11–14]:

- Data cleaning from various noises;
- Elimination of various anomalies in the dataset;
- Removing duplicate data;
• Elimination of missing data;
• Change form of categorical data (data encoding);
• Extracting important data from the dataset;
• Data scaling;
• Dividing the dataset into training and test datasets.

Each of these data preprocessing steps plays an important role in building an artificial intelligence model. In real life, most of the data collected for an AI model fall into the category of categorical data. Therefore, one of the most important steps among these steps is the step of changing the form of categorical data, because most of the other steps cannot be performed without doing this process. In addition, the stage of categorical data encoding is considered one of the most complex stages of data preprocessing [12], because in this process, data that do not have a specific numerical amount are modified to an exact value. As a result of this modification, all categorical data will have a new value represented by numbers. More precisely, each categorical value will have a value that exactly affects the output value of the artificial intelligence model. The complexity of this process is that it is impossible to determine in advance whether the change of form is being carried out correctly. Improper shapeshifting affects the output value of artificial intelligence algorithms incorrectly. As a consequence, artificial intelligence algorithms do not achieve the expected efficiency. That is why the correct implementation of categorical data encoding is one of the current research topics of modern artificial intelligence.

In this research work, the methods of converting the categorical data in the dataset into numerical form, that is, categorical data encoding methods, are studied. A new approach has been developed for choosing the best categorical data encoding method for research. How these methods affect the efficiency of the artificial intelligence algorithm is also studied. In general, the main contribution of this research is as follows:

1. All the research works related to the research of categorical data encoding methods, available in the Google search system and published in journals and various conferences, were analyzed at the same time;
2. A methodology for researching the impact of categorical data encoding methods on artificial intelligence algorithms has been developed;
3. The process of applying 17 categorical data encoding methods to one artificial intelligence algorithm is described (until this paper, the number of most studied methods in one research work did not exceed 10). Therefore, to solve one problem posed in this process, a set of data that has undergone the same initial processing stage is used;
4. An approach to selecting the best method for research from existing categorical data encoding methods has been developed.

In the first step of the research, the literature within the scope of the research is analyzed, and the gaps in the studied studies are identified. Based on the problems identified during the literature analysis, the specific purpose of the research and the problem to be solved are justified. The methodology for solving the problem that is posed in the research is described and the results of experimental tests are presented. At the end of the study, the results of the experiments will be discussed and the general conclusions obtained from the study will be presented.

2. Literature Review

In real life, one major issue is that most of the data collected for solving problems with the help of artificial intelligence are categorical (non-numerical) data [15,16]. That is why many studies have been conducted and are being conducted on research of changing the form in categorical data. One such research work is that of Indian scientists P. Amutha and R. Priya, which can be cited as an example. They researched label encoding and one-hot encoding methods in their study [17]. They studied how label encoding and one-hot encoding methods affect the results of random forest, support vector machine, k-nearest neighbor, naive Bayes, and logistics regression algorithms. Although they actually performed a great study, there are some flaws in their research. The first of these is that all the data in the dataset used in
the study are categorical type. The fact that all data are categorical may not fully reveal the results of the research, because changing the values in all fields to a completely different form can lead to the calculation of completely unexpected values. Second, in this study, only two encoding methods were investigated out of the many methods available. Like these scientists, Andrei Iustin studied the effects of one-hot encoding, target encoding, ordinal encoding, CatBoost encoding, and count encoding methods on the results of linear regression, logistic regression, XGBoost, LightGBM, and random forest algorithms in his master’s thesis [18]. An important achievement of this study is that these methods were evaluated in terms of accuracy and implementation time. Even in this study, a few encoding methods were investigated. In addition, the replacement of missing data with the average value in the field was used. Although the method of eliminating missing data is not very important in this study, it is not recommended to use this method for categorical data. Categorical data were the main object of research and this has a negative impact on research work. Also, in this study, the approach of choosing categorical data encoding methods according to the number of features is proposed, but this approach is not scientifically based. Another similar study was conducted by Aravind Prakash and others [10]. In their research, they used dummy encoding and standard normalization methods to carry out experimental results on categorical data. The main gap in this study was the use of only one categorical data encoding method. At the same time, the results of this study were not sufficiently analyzed, so it remains open how effective this method was. Chinese scientists Chang Liu, Liu Yang, and Jingyi Qular studied how the application of minmax scaling method for numerical data, CatBoost encoding, count encoding, and label encoding methods for categorical data affects the result of neural networks in their research [12]. It is known that the goal of performing the scaling process is to bring the values of all fields to the same range, but this study does not consider it. That is, it was not taken into account that numerical data can be brought to the range [0:1] using the minmax scaling method and that categorical data can exceed this range as a result of using encoding methods. This actually has a negative effect on the results of the experiment. South Korean scientists M. K. Dahouda and I. Joe proposed a new categorical data encoding method based on deep learning in their research [19]. Together with this method, they studied the effects of target, binary, and one-hot encoding methods on logistic regression, multi-layer perceptron, random forest, gradient boosting, and LSTM artificial intelligence models. Although their research work was carried out excellently, some misunderstandings are observed in their results section. The first of these is due to the fact that the results of the experiments conducted on one dataset were not combined. That is, target encoding and binary encoding methods were separately compared, the proposed approach and one-hot encoding were separately compared. The second confound in this study is that although a new method was proposed, one-hot encoding performed well in most of the results. The proposed approach was effective only when used with LSTM. This explains the efficiency of the artificial intelligence model, not the efficiency of the encoding methods. As a result, there were gaps regarding the main result expected from the research. In another study related to this research topic, the authors proposed a two-hot encoding method [20], but the methodology of this method is not sufficiently substantiated and experimental tests are not provided in the research work. In another such study, nine categorical data encoding methods were analyzed [21]. However, the results of research on the impact on artificial intelligence algorithms have not been provided. One-hot encoding and ordinal encoding were reported in another study within this topic [22]. However, experimental tests did not provide accurate information about which of them is better. That is, in the given data, two methods were not used together with one artificial intelligence algorithm. This lead to leaving open the question of which encoding method is better.

In other works in the literature reviewed within the framework of the research, categorical data encoding methods were considered just for the purpose of changing the form. That is, in these studies, which categorical data encoding method is effective has not been studied. In general, based on the analysis of the studied literature, the results presented in Table 1 below were obtained.
Table 1. Analysis of categorical data encoding methods used in articles published within the scope of the research under consideration. “+” represents that the research paper in the row studied the method in the column, the method in the cell with a painted background was found to be the best method in the studied literature; \(a^*\) is the total number of articles in which the method listed in the column was researched, \(b^*\) is the number of times the method listed in the column was recognized as the best method.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>[17]</td>
<td>(\checkmark)</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2023</td>
</tr>
<tr>
<td>[10]</td>
<td></td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2021</td>
</tr>
<tr>
<td>[18]</td>
<td>+</td>
<td>(\checkmark)</td>
<td></td>
<td></td>
<td></td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2023</td>
</tr>
<tr>
<td>[12]</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2020</td>
</tr>
<tr>
<td>[23]</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2019</td>
</tr>
<tr>
<td>[19]</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2021</td>
</tr>
<tr>
<td>[20]</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2021</td>
</tr>
<tr>
<td>[21]</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2022</td>
</tr>
<tr>
<td>[22]</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2022</td>
</tr>
<tr>
<td>[26]</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2022</td>
</tr>
<tr>
<td>[27]</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2018</td>
</tr>
<tr>
<td>[28]</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2017</td>
</tr>
<tr>
<td>[29]</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2020</td>
</tr>
<tr>
<td>[30]</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2018</td>
</tr>
</tbody>
</table>

\(a^*\) 4 8 14 12 3 5 1 0 4 2 2 3 2 4 3 3 3 2

\(b^*\) 1 0 5 0 0 0 0 0 1 0 0 0 1 0 0 0 1 0 -

The research works presented in Table 1 are related to the research of the categorical data encoding methods and include all the research works available in the Google search engine and published in journals and various conferences. As can be seen in Table 1, almost half of the categorical data transformation studies did not provide a conclusion about which encoding method was best for the study. Also, in all the research works where the data encoding method that has the best effect on the research result was found, it was not concluded that only one method is the best. That is, in five research studies, the one-hot encoding method was found to be the best, and in the remaining four cases, four different methods, label encoding, CatBoost encoding, Target encoding, and the polynomial encoding method, were found to be the most effective data encoding method. At the same time, in most of the studied research work, 3–4 or less categorical encoding methods were used as the object of research. Only two research groups, the group of John Hancock and Taghi Khoshgoftaar and the group of M. Ouahi, S. Khoulji, and M.L. Kerkeb studied 10 and 9 methods, respectively [21,29]. However, the same methods have not been studied in these studies. Another conclusion obtained from the analysis of the literature is that interest in this direction of research has increased in recent years. This can be seen from Table 1 and Figure 1.

In general, the following conclusions can be drawn from the studied research works:

- The single categorical data encoding method was not always the most effective for all studies. That is, the determination of the best method should be carried out for each study;
- A clear rule or approach to choosing the best method for research has not been developed;
- There is an urgency to conduct research using all categorical data encoding methods in one dataset.
Figure 1. The number of publications published within the scope of the research under consideration (in terms of years).

Based on these conclusions, selecting the best method of categorical data encoding among all available methods was defined as the research goal. At the same time, it was decided to use the dataset collected to evaluate whether it is possible to allocate credit to customers in order to implement the purpose of the research and to evaluate the results.

3. Materials and Methods
3.1. Steps of Research Implementation

The study of how the application of different methods of transformation of the categorical data in the dataset affects the result of the study requires the implementation of several steps. In a narrow sense, this research work includes the following steps:

Step 1. Determination of dataset for research (choice of dataset containing categorical data);
Step 2. Carrying out preprocessing steps in the dataset (except for the process of categorical data encoding);
Step 3. Analysis categorical data encoding methods;
Step 4. Applying various transformation methods to the categorical data in the collected dataset and saving the results of each transformation;
Step 5. Choosing an intelligent algorithm for evaluating credit allocation to bank customers;
Step 6. Training the selected artificial intelligence algorithm using each dataset saved in step 4;
Step 7. Evaluation of each state of the artificial intelligence model built in step 6;
Step 8. Analysis the evaluation results and determine the best categorical data encoding for the research.

Based on these steps, the general scheme of the research methodology can be described using the processes presented in Figure 2 below.

Figure 2. Scheme of the overview of the research methodology.
Each of the research methodology steps presented in Figure 2 requires specific actions and explanations. Each of these processes is explained in sequence throughout the study.

3.2. Research Materials—Identifying the Dataset for the Study

3.2.1. Dataset Description

This study used a dataset, collected for the purpose of evaluating customer credit allocation, to compare form switching methods in categorical data. This dataset was obtained from kaggle.com [31]. In general, other datasets can be used to conduct this research. However, when selecting a dataset, it is a must to choose one that also contains a categorical data column or columns, representing the research object. This dataset that was used in the research allows for an increase in the comprehensibility of the research together with the fulfillment of the above-mentioned basic conditions. That is, expressing the problem that is posed in the research based on this dataset will be equally understandable to every reader. Also, the categorical data fields in this selected dataset are relevant to the target field value, so these fields cannot be dropped. At the same time, the presence of fields that accept numeric values in the dataset helps to ensure the reliability of the research.

This dataset contains 12 types of data, which are as follows:

- Age of the customer (in integer value);
- Annual income of the customer (in integer value);
- Information on the client’s ownership of the house (in text value);
- The customer’s work experience (in real value);
- The purpose of obtaining a credit (in text value);
- Credit level (in text value);
- Loan amount (in integer value);
- Credit rate—loan percentage (in real value);
- Credit status, 0—non-standard status, 1—standard status (integer value);
- Information about the percentage of the annual income received (in real value);
- Customer’s historical credit status (integer value);
- Duration of the customer’s credit history (in integer value).

It can be seen that three of these fields receive categorical values, and in order to perform actions on them, first of all, it is necessary to convert them to digital form. The first of these fields is the customer’s home ownership field, which accepts one of four values (‘rent’, ‘own’, ‘mortgage’, ‘other’). The second of the text fields is the field representing the purpose of the loan, which accepts the following values: ‘personal’, ‘education’, ‘medical’, ‘venture’, ‘homeimprovement’, ‘debtconsolidation’. The third text value receiving field is the field representing the credit level. It accepts one of the values A, B, C, D, E, F, G, representing 7 levels. The purpose of this research is to find out which form substitution methods will improve the results in the text fields in the 3 fields mentioned above.

3.2.2. Previous Studies Performed on the Dataset

Credit Risk Dataset has been posted on kaggle.com for over 4 years. During this time, several researchers and programmers have implemented their research and projects based on this dataset. Using the Google search engine, works in the literature such as 4 articles and 1 master’s thesis were identified as a result of a search [32–36]. As a result of the analysis of this literature, it was found that work on categorical data encoding was not conducted in these research work. Also, as of 4 August 2024, 68 codes written on the basis of “Credit Risk Dataset” on kaggle.com were studied. In the research, statistical information was collected for the purpose of writing these 68 codes and what categorical data encoding methods were used in them. The results of this statistical analysis can be seen in the diagrams presented in Figures 3 and 4.
It can be seen from Figure 3 that only 4 categorical data encoding methods were used in all the codes written for the Credit Risk Dataset on kaggle.com. In some cases, to be more precise, in 14 of 68 codes, categorical data encoding methods were not used at all. This situation is mainly found in dataset analysis codes. The most used method was the label encoding method, used in 17 codes. The indicators of dummy encoding and one-hot encoding methods were 16 and 15, respectively. In total, in six cases, two categorical data encoding methods were used in one code. However, they are not for the purpose of comparison, it is just that if one field is used, the other is used for another field. More specifically, the codes written for the Credit Risk Dataset at kaggle.com did not analyze categorical encoding methods. This conclusion can also be seen from the statistics of the goals of the codes presented in Figure 4.

It can be seen from Figure 4 that the codes written for the Credit Risk Dataset are mainly focused on 5 goals. These are the following goals:

- Exploratory data analysis;
- Credit risk prediction;
- Credit risk classification;
- Modeling;
- Assessment projects.

In addition, 6 out of 68 codes do not have a specific purpose.

In general, the effect of categorical data encoding methods on artificial intelligence algorithms has not been studied using the Credit Risk Dataset selected in the study.
3.3. Dataset Preprocessing (Except for the Categorical Data Encoding)

In this study, the following preprocessing steps were performed before transferring the data to the artificial intelligence model:

- Elimination of missing data;
- Categorical data encoding;
- Data scaling.

At the stage of elimination of missing values, the process of replacing the missing data in the cell with the most common values in the field is performed. Of course, this method is not the most effective, but since the research did not consider the best method for this stage, this method was simply chosen. This method was used together with all encoding methods.

Obviously, different fields in a dataset will accept different ranges of values. This, in turn, can negatively affect the performance of the artificial intelligence model [37–39]. In order to eliminate this problem, the data scaling process was implemented. In this study, only one method, minmax scaling, was used with all categorical data encoding methods in the data scaling process. That is, all the data in the dataset were scaled using Formula (1) [40,41]:

\[
u_{ij}^{\text{new}} = \frac{u_{ij} - u_{j^{\text{min}}}}{u_{j^{\text{max}}} - u_{j^{\text{min}}}}
\]

where \( u_{ij} \) — the value at the intersection of \( i \)—row and \( j \)—column, \( u_{j^{\text{min}}} \) is the smallest of the values in column \( j \), \( u_{j^{\text{max}}} \) is the largest of the values in column \( j \).

3.4. Categorical Data Encoding Methods

3.4.1. Analysis of Categorical Data Encoding Methods

As mentioned above, there are several fields in the dataset, and the values of these fields are not always represented by numerical values. Non-numerical values hinder the automation of intelligent systems when making decisions or performing other similar actions with the help of artificial intelligence. In this case, discarding fields that do not accept a numeric value can have a very negative effect on the results of artificial intelligence. Therefore, it is necessary to transfer categorical data to a numerical value at the stage of preprocessing data. Before considering the methods of implementation of this process, it is necessary to know the following two kinds of categorical data:

- Ordinal data;
- Nominal data.

Ordinal data are textual data that indicate the level. The following data groups can be examples of ordinal data:

- Few, average, many;
- Very unhappy, unhappy, good, happy, very happy;
- Very long, long, medium, short, very short,

Nominal data are textual data that do not represent any semantic level. Examples of nominal data can include the following data groups:

- Blue, white, light blue, blue, etc.;
- Samarkand, Tashkent, Jizzakh, Navoi, Bukhara, Khorezm, etc. (cities of Uzbekistan);
- Cows, sheep, horses, dogs, cats, goats, donkeys, chickens, etc.

In the course of this study, the following most common methods of categorical data encoding are analyzed in sequence:

- Label encoding;
- Ordinal encoding;
- One-hot encoding;
- Dummy encoding;
- Effect encoding (deviation encoding, sum encoding);
- Binary encoding;
• Base N encoding;
• Gray encoding;
• Target encoding (mean encoding);
• Count encoding;
• Frequency encoding;
• Leave-one-out encoding;
• CatBoost encoding;
• Hash encoding;
• Backward difference encoding;
• Helmert encoding;
• Polynomial encoding.

It should be emphasized that these categorical data encoding methods are the methods proposed in the studied literature and on the sites of the open Internet system. That is, in the study, these methods were not chosen according to any particular feature, but all the methods of the studied research work were combined in this research work.

Label encoding: In this categorical data encoding method, each unique textual value is shifted to an integer value based on the sequence. That is, if the set of unique values in a field consisting of n lines is \( A = \{a_0, a_1, a_2, \ldots, a_{k-1}\} \), then the form substitution in text data in this field (2) is achieved using the following formula:

\[
\text{TEnc}(b_i) = \begin{cases} 
0, & b_i = a_0; \\
1, & b_i = a_1; \\
2, & b_i = a_2; \\
\vdots & \\
k - 1, & b_i = a_{k-1}; 
\end{cases}
\]  

(2)

Here \( i = 1 \ldots n \). Although label encoding is considered the simplest and most understandable method of categorical data encoding, this method is usually used when the categorical data in one field have only two unique values. That is, this method is more effective only in cases where \( k = 2 \), because in the label encoding method, text values are replaced by numeric values without paying attention to any level. As a result, the values set by the label encoding are misinterpreted by machine learning algorithms as having mathematical significance, and the model’s accuracy index drops. Since the number of unique values of the majority categorical data in a data field in real life is more than two, it is difficult to call this method widespread.

Ordinal encoding: This method is similar to the label encoding method. The main difference between them is that in the label encoding method, the replacement of integers with text data is performed according to their sequence, while in the ordinal encoding method, they are replaced according to their semantic compatibility. That is, in the method of ordinal encoding, categorical data with a high semantic value are replaced with a larger integer, and categorical data with a low semantic value are replaced with a smaller integer. It follows that applying this method only to ordinal data helps to achieve accuracy. It should be mentioned that in the ordinal encoding method, the data must first be categorized according to the levels. A unique value must not be left in the field, otherwise an error will occur. In order to avoid this, the unique values of the field to be changed should be determined.

It can be seen from Figure 5 that the set of unique values \( A \) in the rating field is defined as follows: \( A = \{"excellent","sufficient","good","insufficient"\} \). These values are semantically reordered in the following sequence: \( A' = \{"insufficient","sufficient","good","excellent"\} \) and they are respectively \( \{0, 1, 2, 3\} \) changed to numeric values.
One-hot encoding: Applying the above two methods (label encoding and ordinal encoding) for nominal data may not be effective. Because these methods, as mentioned above, give numerical values of different magnitudes to nominal values of the same level, this can lead to errors. In this case, one-hot encoding is considered as an effective method. In this method, each unique value is added to the dataset as a separate field. That is, if there are \( k \) unique values in one text field in the dataset, \( k \) new columns are added to the dataset. Cells \( (b_{i,j}) \) in these columns are filled with 0 or 1. Taking the value 0 or 1 is determined by Formula (3) as follows:

\[
b_{i,j} = \begin{cases} 
0, & b_{i,j} \neq a_j; \\
1, & b_{i,j} = a_j 
\end{cases} 
\]

(3)

here \( a_j \)—unique values in a text field; \( j = 0 \ldots k - 1; i = 1 \ldots n \).

The disadvantage of the one-hot encoding method is that if there are a lot of unique values in the field where the form substitution is being performed, it will automatically increase the size of the data. Therefore, this method is recommended for cases where there are not many unique values.

Dummy encoding: This method is very similar to the one-hot encoding method. The main difference between these two methods is that one-hot encoding generates \( k \) fields for \( k \) unique values, while dummy encoding generates \( k - 1 \) fields for \( k \) unique values. That is, in the dummy encoding method, the first unique value line is filled with only 0 values. The values in cells \( (b_{i,j}) \) of fields created as a result of dummy encoding are created by Formula (4) as follows:

\[
b_{i,j} = \begin{cases} 
0, & b_{i,j} \neq a_j; \\
1, & b_{i,j} = a_j 
\end{cases} 
\]

(4)

here \( j = 1 \ldots k - 1; i = 1 \ldots n \).

Overall, the dummy encoding method is a small improvement over the one-hot encoding method.

Effect encoding (deviation encoding, sum encoding): This method is almost the same as the dummy encoding method. The main difference is that in the effect encoding method, as in the dummy encoding method, one line is filled with the value \(-1\) instead of 0 (5).

\[
b_{i,j} = \begin{cases} 
-1, & b_{i,j} = a_0; \\
0, & b_{i,j} \neq a_j; \\
1, & b_{i,j} = a_j 
\end{cases} 
\]

(5)

Binary encoding: This method is similar to the combination of one-hot encoding and label encoding methods. That is, in the binary encoding method, unique values are first matched with sequential numbers in the binary numbering system, as in the label encoding method. In the next step, the length of the binary number with the maximum length \((l)\) is determined. The binary values of the remaining unique values are extended to \( l \) length (by adding 0 s in front of the binary values). In the next step, \( l \) fields are created and the
bits of the binary values of each unique value are sequentially placed in the fields. As a result, the values in the text field are replaced by \( l \) fields that accept combinations of 0 s and 1 s, as in the one-hot encoding method. In the one-hot encoding method, \( k \) fields are created for \( k \) unique values, while in the binary encoding method, \( l \) new fields are created. For comparison, \( l \) can be found by Formula (6) as follows:

\[
l = \begin{cases} 
\lfloor \log_2 k \rfloor, & \text{if } \lfloor \log_2 k \rfloor = \log_2 k; \\
\lfloor \log_2 k \rfloor + 1, & \text{if } \lfloor \log_2 k \rfloor < \log_2 k
\end{cases}
\]  

(6)

where \([x]\) is the notation representing the integer part of the number \(x\).

Base N encoding: The base N encoding method is an extended version of the binary encoding method. That is, in the base N encoding method, it is possible to change the form of text data based not only on the binary numbering system, but also on other existing \(N\)-based numbering systems. In this method, the larger the base of the number system, the smaller the number of fields \(l'\) will be after applying the method. It can also be seen from Formula (7) for determining \(l'\):

\[
l' = \begin{cases} 
\lfloor \log_N k \rfloor, & \text{if } \lfloor \log_N k \rfloor = \log_N k; \\
\lfloor \log_N k \rfloor + 1, & \text{if } \lfloor \log_N k \rfloor < \log_N k
\end{cases}
\]  

(7)

Gray encoding: The gray encoding method can be said to be a form of binary encoding method. The main difference of this method from the binary encoding method is that each sequence of unique values differs from the binary value of the preceding and following unique value by only one bit (Figure 6).

![Figure 6. The result of using the gray encoding method (grade field). * in alphabetical order.](image)

As the number of unique values increases, changing only one bit of the sequence reduces miscoding errors and provides more flexibility in terms of synchronization.

Target encoding (mean encoding). The target encoding method is a method of replacing the field values with textual information with the average value of the corresponding values of the field representing the target. That is, in the target encoding method, each unique value \(b_i\) is replaced by the average value of the values in the intersection of the lines belonging to this unique value with the target area (8) [40]:

\[
TEnc(b_i) = \frac{\sum_{j=1} t_{ij}}{\sum_{j=1} f_j}
\]  

(8)

where \(t_{ij}\) is the target field values of the unique value \(b_i\).

Count encoding: In the count encoding method, each unique value in a text field is replaced by the number of occurrences of that unique value. From this definition and Formula (7), the form substitution in text data using the count encoding method is carried out by Formula (9) as follows:

\[
TEnc(b_i) = \sum_{j=1} j
\]  

(9)
where $\sum_{j=1}^j$ is the total number of cells in the text field whose value is equal to $b_i$.

Frequency encoding: The frequency encoding method is almost the same as the count encoding method, only, in this method, the textual data are exchanged not by the number of repetitions, but by the frequency of repetitions (10).

$$TEnc(b_i) = \frac{\sum_{j=1}^j}{\sum_{i=1}^i j}$$

Leaves-one-out encoding: This method is similar to the target encoding method. The main difference is that each unique value is not directly replaced by the average value of the target field values corresponding to unique values of that type. When calculating the average value, the value at the intersection of the row, being filled, and the target field is not added to the total sum (11).

$$TEnc(b_i) = \frac{\sum_{j=1}^j t_{ij} - t_i}{\sum_{j=1}^j j}$$

where $t_{ij}$ is the value at the intersection of the $ith$ line and the target field.

CatBoost encoding: The CatBoost encoding method is also a method that serves to change the form of text fields belonging to the target encoding family. In this method, each unique value is replaced by the average of the average values of the previous unique values in the target field and the average of the values of the total in the target field (12).

$$TEnc(b_i) = \begin{cases} \frac{\sum_{j=1}^j t_{ij}}{\sum_{j=1}^j j}, & b_i \neq b_e \\ \frac{\sum_{j=1}^j t_{ij} + \sum_{h=1}^h t_{ih}}{1 + \sum_{h=1}^h h}, & b_i = b_e \end{cases}$$

where $i = 1 \ldots n, e = 1 \ldots i - 1, \sum_{h=1}^h h$ is the total number of $e$’s satisfying the condition $b_i = b_e$ in the text field, $\sum_{h=1}^h t_{ih}$—the sum of values in the target field of lines satisfying the condition $b_i = b_e$ in the text field.

Hash encoding: The hash encoding method implements form substitution in categorical data based on hash functions (md5, sha256, sha512, sha1, blake2b, sha224) studied in information security and database sciences [30,31]. In this method, as in the one-hot encoding method, new fields are created as a result of encoding. The number of these fields is equal to the length of the result of the selected hash function. For example, if encoding is performed using the md5 hash function, the number of new fields will be 512.

Backward difference encoding: This method is similar to a combination of the dummy encoding and ordinal encoding methods. In this method, new fields such as dummy encoding are created and the lines corresponding to the first categorical data are filled with 0, the first column of the lines corresponding to the second categorical data, the first and second columns of the lines corresponding to the third categorical data, and so on, the first, second, . . . , fill the $k - 1th$ columns with the number 1 and the remaining columns with 0. The similarity of this method to ordinal encoding is that the result of this encoding helps to understand the additional effect of each level compared to the previous level. That is, it is recommended to use this method when the sequence of categorical data levels is clear.

Helmert encoding: This encoding method is similar to the backward difference encoding method. They differ only in the encoding process by the value in one column. That is, in this method, the $i - k$ value of the $ith$ column in the line corresponding to the $ith$ categorical data is set, and the value in all other columns is the same as the value in the backward difference encoding method.

Polynomial encoding: In the polynomial encoding method, $k-1$ new columns are created. This method is used to determine polynomial trends (linear, quadratic, cubic, etc.) of ordered categorical variables with respect to the target area [28].
3.4.2. An Approach to Selecting the Best Method for Research

Above, it was demonstrated that there are several types of methods of changing the form of categorical data. From the results of the analysis, it can be seen that some of these methods have mutual similarities. Taking these similarities into account, the 17 categorical data encoding methods studied above can be conditionally divided into 3 classes. These are the following:

- **Class A**: label encoding, ordinal encoding, backward difference encoding, Helmert encoding, and polynomial encoding;
- **Class B**: one-hot encoding, dummy encoding, effect encoding, binary encoding, base N encoding, gray encoding, hash encoding;
- **Class C**: target encoding, count encoding, frequency encoding, leave-one-out encoding, CatBoost encoding.

It should be noted that, in general, there are no clear rules about which of these methods should be used for research. However, based on the results of the analysis, it can be said that attention must be paid to some aspects when choosing methods. These aspects are:

- Whether categorical data are nominal or ordinal;
- The number of unique values in the categorical data field;
- Existence of the hypothesis that the number of repetitions of unique values in the categorical data field is significant for the target field;
- Limited computing time and memory space.

Considering these aspects in the study, the approach shown in Figure 7 below is proposed to choose the best method for research from categorical data encoding methods.

**Figure 7.** An approach to selecting the most appropriate method of categorical data encoding.

In the proposed approach, condition 1 checks whether the data being transformed are of ordinal type. If these data are ordinal, condition 2 is checked, otherwise condition 3 is checked. Condition 2 checks if the number of unique values in the field being transformed is greater than 2. If this condition is met, condition 4 is checked, otherwise it suggests the label encoding method as the best categorical data encoding method. Condition 3 checks the hypothesis that the number of occurrences of unique values in a field is correlated with the target field in the dataset. If condition 3 is true, it is recommended to select one of the methods in Class C as the most suitable categorical data encoding method; otherwise, choose one of the methods in Class B. Condition 4 checks whether there is a clear-fixed sequence of data considered as ordinal. That is, in some cases, it is difficult to determine a strict sequence in an ordinal dataset. For example, in the dataset used in the study, the values in the field providing information on the customer’s homeownership are ordinal, but the order of these values is unclear. That is, there is confusion in determining the
order of the values ‘rent’, ‘mortgage’, and ‘other’ in this field. Thus, if condition 4 is true, one of the methods in Class A is proposed as the best categorical data encoding method. Otherwise, it is recommended to choose one of the methods in Class B.

Of course, choosing one of methods in Class A, B, and C also has its own rules. If condition 4 takes a true value, it is necessary to choose one of the ordinal encoding, backward difference encoding, Helmert encoding or polynomial encoding methods as the best method. In this case, if it is taken into account that ordinal encoding does not create a new field as a result of the change of form, it gives a good result when the number of unique values is large. If it is estimated that there is a linear or quadratic or cubic and inverse relationship between the unique values, the polynomial encoding method is chosen. Backward difference encoding and Helmert encoding methods have almost the same transformation process. However, the Helmert encoding method is more flexible than the backward difference encoding method.

All methods in Class B require the creation of new fields. If there are memory and time limits, one-hot encoding and hash encoding methods will not meet the requirements. The best method in this case may be the base N encoding method. However, this method gives different values to various new values with the same meaning, such as the label encoding method. It could potentially lead to an increase in errors. Therefore, one of either binary encoding or gray encoding can be chosen as a reliable method in this case. The better of these two methods is the gray encoding method, because it reduces miscoding errors due to only one bit of the sequence changing as the number of unique values increases and provides more flexibility in terms of synchronization.

When choosing one of the methods of Class C, firstly, two features should be given attention. One of them is the validity of the assumption that the values of the unique values in the categorical data field depend on the number of repetitions. If this assumption is correct, one of the count encoding or frequency encoding methods can be chosen. The second feature is the assumption that unique values in the categorical data field depend on the sum of values in the target field. If this hypothesis is true, one of the target encoding, leave-one-out Encoding, or CatBoost encoding methods can be used as a best method. Here, one more condition can be added to choose the best method in a more precise way. That is, if the values of the same unique values that are used for form substitution should be the same, the target encoding method can be chosen as the best categorical data encoding method.

3.5. Model of Artificial Intelligence

It is appropriate to use unsupervised learning models of artificial intelligence to classify bank customers based on the dataset used in the research. It is known that there are several artificial intelligence models that allow for dividing objects into classes [42–45]. However, since the main goal of this study is to research categorical data encoding methods, it is enough to choose one model. That is, it is not necessary to research several classification algorithms and find the one with the highest accuracy. It is enough to choose one of the classification models that allows you to solve the problem. In this study, the simplest and most understandable logistic classification model is used in order to divide bank clients into groups for lending. It is known that this classification model is based on the logistic (sigmoid) Function (13).

\[
\sigma(z) = \frac{L}{1 + e^{-k(z-z_0)}}
\]  

(13)

where \(L\) is the supremum of the function; \(z_0\) is the central point of the function;

\(k\) is the coefficient determining the steepness of the graph of the function (the growth rate of the function).

3.6. Evaluation of Artificial Intelligence Model

It is known that one of the most important stages in any research is the process of evaluating the research results. There are several evaluation criteria used in AI classification problems. The most commonly used of them are accuracy, precision, recall, F-1 score. Of
course, there are other evaluation methods, but they do not fully represent how well the classification model works.

The process of evaluating artificial intelligence using accuracy, precision, recall, and F-1 score methods is closely related to the following concepts [46,47]:

- **True positive (TP)**—the given observation is positive and the predicted value is also positive;
- **False positive (FP)**—the given observation is negative, but the predicted value is positive;
- **True negative (TN)**—the given observation is true and the predicted value is negative;
- **False negative (FN)**—a given observation is actually positive but is predicted to be negative.

Accuracy (A) is the simplest and most easily understood evaluation metric and measures the proportion of correctly found observations (14).

\[
A = \frac{TP + TN}{N} \tag{14}
\]

Here, \(N\) is the total number of observations used in the estimation, \(N = TP + TN + FP + FN\). It is recommended to use this estimation method for a balanced data-set. That is, for unbalanced datasets, this method may misclassify model accuracy.

For unbalanced data, it is appropriate to use the precision \((P)\) and recall \((R)\) assessment methods. Precision is used to represent the proportion of cases that are predicted to be positive and are actually positive (15).

\[
P = \frac{TP}{TP + FP} \tag{15}
\]

Recall is the percentage of true positive cases that are marked as positive. That is, it measures what percentage of true positives are assumed to be positive (16).

\[
R = \frac{TP}{TP + FN} \tag{16}
\]

The F-1 score is a harmonic mean of precision and recall that can be applied to both balanced and unbalanced datasets (17).

\[
F - 1\text{Score} = \frac{2 \cdot P \cdot R}{P + R} \tag{17}
\]

### 3.7. Implementation

Python programming language was used to implement the proposed methodology in the study. The following python libraries were used:

- **Pandas**—to perform actions on data downloaded from Kaggle;
- **Scikit-learn**:
  - sklearn.impute module for missing data imputation (SimpleImputer);
  - sklearn.preprocessing module for data scaling (MinMaxScaler) and categorical data encoding;
  - sklearn.model_selection module for data collection for splitting training and test datasets (train_test_split);
  - sklearn.pipeline module for pipeline construction (Pipeline);
  - sklearn.linear_model module for artificial intelligence model (LogisticRegression);
  - sklearn.metrics module for artificial intelligence model evaluation for (accuracy_score, recall_score, precision_score, f1_score);
- **category_encoders**—to use categorical data encoding methods that are not available in the sklearn.preprocessing module.

### 4. Results

In the study, the dataset “Credit Risk Dataset” obtained from kaggle.com was used for the purpose of researching the categorical data encoding methods [31]. In the research, a data-set of 4034 records included in this dataset is divided into training and test sets.
in the ratio of 20:80, and experimental tests are conducted. The dataset went through a preprocessing step before being passed to the AI algorithm. Missing data were replaced with the most repeated values in the field. The minmax scaling method was used to scale the data. In order to convert 3 categorical data into numerical form, the above 17 categorical data encoding methods were used to transform the form. After each method, an artificial intelligence model was trained and tested using an artificial intelligence algorithm based on a logistic function. Test results were evaluated using accuracy, precision, recall, and F-1 score evaluation methods. The results of this experiment are shown in Table 2 and Figure 8.

Table 2. The effect of the use of different categorical data encoding methods on the result of the artificial intelligence algorithm (the results are sorted by F1_score in descending order).

<table>
<thead>
<tr>
<th>№</th>
<th>Encoding Methods</th>
<th>Accuracy</th>
<th>Evaluation Methods</th>
<th>F1_Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Gray encoding</td>
<td>0.8203</td>
<td>Recall</td>
<td>0.3991</td>
</tr>
<tr>
<td>2</td>
<td>Sum encoding</td>
<td>0.8166</td>
<td>Precision</td>
<td>0.6551</td>
</tr>
<tr>
<td>3</td>
<td>One-hot encoding</td>
<td>0.8153</td>
<td></td>
<td>0.7066</td>
</tr>
<tr>
<td>4</td>
<td>Helmer encoding</td>
<td>0.8153</td>
<td></td>
<td>0.754</td>
</tr>
<tr>
<td>5</td>
<td>Polynomial encoding</td>
<td>0.8153</td>
<td></td>
<td>0.7517</td>
</tr>
<tr>
<td>6</td>
<td>Dummy encoding</td>
<td>0.8111</td>
<td></td>
<td>0.4407</td>
</tr>
<tr>
<td>7</td>
<td>Backward difference encoding</td>
<td>0.5481</td>
<td>Recall</td>
<td>0.436</td>
</tr>
<tr>
<td>8</td>
<td>Leave-one-out encoding</td>
<td>0.5438</td>
<td>Precision</td>
<td>0.4265</td>
</tr>
<tr>
<td>9</td>
<td>CatBoost encoding</td>
<td>0.5438</td>
<td></td>
<td>0.7586</td>
</tr>
<tr>
<td>10</td>
<td>Base N encoding (base = 3)</td>
<td>0.5438</td>
<td>Recall</td>
<td>0.417</td>
</tr>
<tr>
<td>11</td>
<td>Count encoding</td>
<td>0.5438</td>
<td>Precision</td>
<td>0.417</td>
</tr>
<tr>
<td>12</td>
<td>Frequency encoding</td>
<td>0.5438</td>
<td></td>
<td>0.417</td>
</tr>
<tr>
<td>13</td>
<td>CatBoost encoding</td>
<td>0.5438</td>
<td></td>
<td>0.7586</td>
</tr>
<tr>
<td>14</td>
<td>Ordinal encoding</td>
<td>0.5438</td>
<td></td>
<td>0.7586</td>
</tr>
<tr>
<td>15</td>
<td>Binary encoding</td>
<td>0.5438</td>
<td></td>
<td>0.7586</td>
</tr>
<tr>
<td>16</td>
<td>Hashing encoding</td>
<td>0.5438</td>
<td></td>
<td>0.7586</td>
</tr>
<tr>
<td>17</td>
<td>Label encoding</td>
<td>0.5438</td>
<td></td>
<td>0.7586</td>
</tr>
</tbody>
</table>

Figure 8. The effect of the use of different categorical data encoding methods on the result of the artificial intelligence algorithm. In this figure “Backward Difference…” is Backward difference encoding.
As can be seen from Table 2 and Figure 8, categorical data encoding methods have influenced the result of the artificial intelligence model. The categorical data encoding methods that recorded the best three results belong to Class B. Among them, the best result was observed by the gray encoding method, with values for accuracy, precision, recall, and F-1 score of 0.8203, 0.4454, 0.7704, and 0.5645, respectively. The lowest indicator was observed in the label encoding method, which recorded 0.7719, 0.2701, 0.6551, and 0.3825 indicators, accordingly. It can be said that the main reason for this is that there are more than two unique values in the fields where the form is being changed. Another categorical data encoding method with low results was the hashing encoding method. Its evaluation indicators were 0.7980, 0.3696, 0.7222, and 0.4890, respectively. In the rest of the methods, the results were almost identical. Another important aspect of the results of the study is that the results of C-class methods occupy the middle places. This means that the implementation process of these methods is very close together.

The results were obtained using the methodology developed in the study. In other words, the most effective method for encoding categorical data was determined by testing all 17 methods. It should be noted that if the suggested approach had been used in the research instead of this method, the methods of Class B would have been chosen. That is, one of the methods that showed the highest result according to the results of the experiment would be chosen. That is, according to the conditions of the approach to the selection of the best method, the dataset in the study would have suggested the choice of one of the methods from Class B. From the result, it can be seen that the methods with the first three best indicators belong to Class b. This proves that the developed approach is reliable.

5. Discussion

5.1. Research Achievements

It can be seen from the results of the research that although the performance of the artificial intelligence model is not high, the goal of the research has been achieved. That is, the aim of the study was to study the impact of the methods of changing the form on textual data on the models of artificial intelligence, and this goal was fulfilled. This can be seen in the difference between the best and worst indicators. That is, the results of using the gray encoding method improved by 0.0484, 0.1753, 0.1153, and 0.1820, compared to the results of using the label encoding method. More specifically, as a result of this study, the accuracy indicator improved by 4.84%, the precision indicator by 17.53%, the recall indicator by 11.53%, and the F-1 score indicator by 18.20%. These changes are considered a good indicator for any study and indicate that the results of this study are positive.

Another achievement of this study was that based on the analysis of 14 studied methods, an approach was developed to determine the best categorical data encoding method for research. Although the developed approach cannot be completely automated, it simplifies the workflow of the artificial intelligence engineer. It should be noted that the processes performed based on the methodology developed in the study (methodology for researching the impact of categorical data encoding methods on artificial intelligence algorithms) were fully automated. Experimental tests were carried out and results were obtained based on this automated artificial intelligence system.

5.2. The Superiority of the Research Results over the Results of Previous Studies

The main difference of this study from the previous studies conducted within this topic is that 17 categorical data encoding methods were studied and their results were analyzed in the study under consideration. In previous studies, at most, 10 categorical data encoding methods were studied [29]. Also, this study does not provide a comparative analysis of the impact of methods on artificial intelligence models. The one-hot encoding method, which was considered the best method in most of the studies prior to this study, ranked third in terms of accuracy, recall, and F-1 score evaluation indicators in the study under consideration. Precision took the eighth place according to the evaluation indicator.
Another important difference between the results of the research and the results of the previous research is the presence of an approach to choosing the best categorical data encoding method. Until now, only one study provided partial guidelines for choosing the best categorical data encoding method [18], but in this literature, there is only a brief opinion on which of the five methods to choose. In this study, an approach for selecting the best categorical data encoding method from 17 methods is proposed. Instructions for implementing this approach are also provided.

5.3. Limitations

From the results of the experiment, it can be seen that the gray encoding method returned the best result for all four evaluation methods. However, these results may not satisfy those interested in using the artificial intelligence model. It should be noted that the results of this evaluation might not satisfy the artificial intelligence model in general. That is, the F-1 score evaluation index of 0.5645 does not represent high accuracy. The main reasons why the indicators are not high in this study are as follows:

- Selection of optional and simplest methods at the stage of initial data processing (elimination of missing data and scaling);
- Voluntary selection of the artificial intelligence model;
- The selected artificial intelligence model is trained and tested based on standard parameters.

Perfecting these steps requires additional time and research. Of course, performing these actions was not chosen as an important part of the study, but it is one of the limitations of this study.

Also, one of the main limitations of the research is that there is no possibility of fully automating the approach of choosing the best categorical data encoding method. In order to fully automate this process, it is necessary to determine the hypotheses related to categorical data. This process is currently carried out in connection with a person.

6. Conclusions

As a general conclusion from the reviewed research, the correct selection of categorical data encoding methods was considered one of the most important stages in the design of an artificial intelligence model. It became known from the studied literature that there are several ways of changing the form of textual information. But the main complexity of the research is that the best categorical data encoding method changes depending on the types of data used to solve the problem. That is, it is considered that the method of changing the form is not the best for all studies in clear single-text data, the best method should be determined for each study. Therefore, in this research work, a methodology for determining the best method of changing the form of textual information was developed and experiments were conducted based on the developed methodology. During the experiment, 14 categorical data encoding methods were studied. Using these methods, an artificial intelligence model was built to assist customers in making decisions on credit. This artificial intelligence model was evaluated by accuracy, precision, recall, and F-1 score. The evaluation results show that the application of categorical data encoding methods to the artificial intelligence model has a clear effect on the overall results. This can be seen in the differences between the best and the lowest values returned in the experimental results. As a result of the experiments in this study, the accuracy of the artificial intelligence model improved by 4.84%, the precision by 17.53%, the recall by 11.53%, and the F-1 score by 18.20%. In summary, the intended purpose of the research was achieved.

Of course, it was determined that the results of this study have limitations within their scope. The first of these shortcomings is the optimization of some steps of the methodology, which is considered in detail in the research. It is possible to carry out this process in general. One of them is to combine this research with the codes on the dataset used in the research on kaggle.com. The second limitation is to automate the approach of choosing the best categorical data encoding method. Although there are currently no clear solutions to this shortcoming, future research work will be carried out to eliminate both limitations.
Despite the limitations, the application of the methodology and approach developed in this study to many artificial intelligence projects is effective. More precisely, in all artificial intelligence models that contain categorical data, categorical data can be used to convert categorical data into a digital representation. As a result, at least one problem of the artificial intelligence engineer in building an optimal artificial intelligence model is eliminated and time efficiency will be achieved.

**Author Contributions:** This article was written and designed by F.B., R.N., A.R. and F.A. who analyzed and validated the proposed model. Y.-I.C. supervised the study and contributed to the analysis and discussion of the algorithm and experimental results. All authors have read and agreed to the published version of the manuscript.

**Funding:** This study was funded by the Korea Agency for Technology and Standards in 2022. The project numbers are 1415181629 (Development of International Standard Technologies based on AI Model Lightweighting Technologies), 1415180835 (Development of International Standard Technologies based on AI Learning and Inference Technologies). Also this research was supported by the Gachon University research fund of 2024 (GCU-202400560001).

**Data Availability Statement:** Dataset: https://www.kaggle.com/datasets/laotse/credit-risk-dataset/data (accessed on 4 August 2024); Code: https://colab.research.google.com/drive/10SYb9Ju_uDxR00H-auFsFjGeXVUm4?usp=drive_link (accessed on 4 August 2024).

**Acknowledgments:** The authors would like to express their sincere gratitude and appreciation to Young-Im Cho (Gachon University) for her support, comments, remarks, and engagement over the period in which this manuscript was written. Moreover, the authors would like to thank the editor and anonymous referees for the constructive comments in improving the contents and presentation of this paper.

**Conflicts of Interest:** The authors declare no conflicts of interest.

**References**


**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.