International Journal of Engineering Trends and Technology
ISSN: 2231–5381 / https://doi.org/10.14445/22315381/IJETT-V72I8P123

Volume 72 Issue 8, 237-243, August 2024

Original Article

Ranking Learning Algorithm for Likert Scale (RLALS) for Prediction of Student Perceptions about Curriculum, Teaching-Learning and Research

Dharmendra Patel[1], Sanskruti Patel[2], Atul Patel[3], Jay Nanavati[4], Unnati Patel[5], Ankit Faldu[6], Amisha Shingala[7]

[1,2,3,4,5,6,7]*Smt. Chandaben Mohanbhai Patel Institute of Computer Applications, Gujarat, India.*

[1]*Corresponding Author : dharmendraptel.mca@charusat.ac.in*

Received: 20 April 2024 Revised: 12 July 2024 Accepted: 31 July 2024 Published: 28 August 2024

Abstract *- The Likert scale is an important aspect of collecting data for various situations. The data is in ordinal form, so performing analysis and prediction requires a special kind of algorithm. In this paper, a Ranking Learning Algorithm for the Likert Scale (RLALS) is proposed to predict ordinal data. The data from the education domain is collected for experimentation. The data related to the feedback process in the context of curriculum, teaching-learning, and research was collected from 339 students with 12 different parameters. The proposed algorithm is compared with a well-known logistic regression model. The accuracy of the proposed model before feature engineering and after feature engineering is better than logistic regression. The accuracy of the proposed model before feature engineering is 68.63%, while after feature engineering, it is 89.24%.*

Keywords *- Logistic regression, Likert scale, Ordinal data, Simple linear regression, p-value.*

1. Introduction

The foremost viable tool we have for illuminating ourselves and the world is education. Typically, genuine since a quality education gives one the abilities vital to correctly comprehend data and apply it to circumstances experienced in the way of life [1]. The curriculum has drawn more consideration in higher instruction over the past ten a long time. The curriculum is one of the key concepts in a talk on higher instruction. The word "curriculum" has a few diverse implications. The ethical, political, and ideological objectives basic to the numerous conceptualizations of education have been examined for decades in curriculum thinking [2]. Students are the spine of any organization, so their perceptions of curriculum, teaching, learning, and research are essential for the growth of any institution. Education organizations which accept quality take students' perceptions into account in curriculum, teaching, learning and research. Suppose any organisation is able to predict the perceptions of students based on historical data. In that case, it helps the organisation plan their activities in different aspects in a well-planned manner. Generally, the perceptions of students are gathered on a Likert scale [6-8]. Definition 1.1: The Likert scale consists of n unidimensional scales {s1, s2,... sn} to collect respondents' perceptions about certain situations. Such a scale has several practical benefits; it is practical, simple, and seems natural to use for many applications. However, the prediction of the Likert scale class needs a specialized type of multi-classification algorithm. The main objective of the paper is to predict the perceptions of students, which are in ordinal data form, with the proposed ranking learning algorithm. Section II of the paper will discuss the Likert scale and data analysis. Section III will discuss the well-known Logistic Regression Algorithm. Section IV will give details about the proposed ranking learning algorithm for the Likert scale. Section V will discuss the dataset, experiments, and result analysis. Finally, the paper will end with conclusions.

2. Related Work

Bayesian methods are extensively used for categorical data analysis in the literature [9-12]. The beginning of categorical data analysis was done with a uniform prior distribution for the binomial parameter. The sample proportions for contingency tables are standard Maximum Likelihood (ML) estimators of multinomial cell probabilities. Insufficient data may have unwanted characteristics. Log-normal and gamma priors in estimating association factors in contingency tables were introduced to lessen the limitations of the above approach. Bayesian methods for estimating multinomial probabilities in contingency tables using a Dirichlet prior distribution were introduced in 1965. Leonard and Hsu (1994) [13] selectively examined the development of Bayesian approaches to the analysis of categorical data. There are two types of categorical data analysis strategies: those that focus on modelling [14, 15] and those that focus on hypothesis testing [16, 17]. Modelling generally focuses on the analysis of categorical data and tries to identify meaningful insights from that. However, the accuracy of the model is not good in comparison to numerical data. The following table describes the gist of research in the analysis of categorical data. Overall, the existing methodologies for categorical data analysis have certain limitations and do not give the appropriate accuracy, particularly when the categorical data have some intrinsic order to them. The majority of methodologies are complex and suitable for numerical data.

3. Research Gaps of Existing Work

- Bayesian methods are used extensively for categorical data analysis. However, it suffers from complexity issues, so they depends on optimization techniques heavily. It entirely depends on posterior distribution, and for that, prior is required. Selection of prior is very difficult in this approach.
- Regression modelling is another vital approach for categorical data and is used extensively in literature surveys. It gives good results if the dataset is small and the veracity of the data is good. It is not suitable for large datasets, and data is nonlinear in nature. It gives good results for numerical data. For the categorical data, the result is not promising.
- Logistic regression modelling is best suited for categorical data. The major problem is of overfitting of the data. In reality, data is not available in linearly separable form. It is based on the assumption that all independent variables must be related linearly with dependent variables.
- Hypothesis testing is one way to cater for categorical data. However, this technique is used primarily for testing hypotheses.
- Overall, Logistic regression is well known and handles categorical data efficiently, but it has an overfitting problem and all independent variables must be linearly related with dependent variables.

4. The Likert Scale and Data Analysis

In 1932, the Likert scale was made as a 5-point scale, which is still broadly utilized nowadays. The foremost point-by-point scales, which require respondents to specify their level of agreement, approval, or belief, range from a collection of common issues. The Likert scale has a few vital qualities, such as: No matter how obvious the association between a thing and a sentence could seem, the things ought to be straightforward to relate to the answers. It is important to note that indeed in spite of the fact that a Likert scale with 5 things is the foremost ordinary, utilizing more things increments the precision of the results. Likert and others, for the most part, agree that it is ideal to receive a scale that's as wide as conceivable. The answers can continuously be condensed into sensible groupings on the off chance that they are essential for examination. There are four fundamental essential levels of estimation of information: nominal, ordinal, interval, and ratio [18]. Ordinal data is used to examine data, especially when it comes to survey scales like Likert or others. The analysis of Likert scale data is done with both descriptive and inferential statistics. There are three main techniques of descriptive statistics that deal with such data effectively: frequency distribution, central tendency, and measure of variability, particularly range. As far as inferential statistics are concerned, several non-parametric tests deal with ordinal data effectively for hypothesis testing. The only limitation of Likert scale data analysis is that it violates the assumptions of parametric tests.

Table 1. Gist of literature survey for the analysis of categorical data

Approach	Descriptions
Bayesian methods	• It is used widely for categorical data analysis • Selection of prior is difficult in this approach • Complexity is more so optimization is required • Posterior distribution is heavily dependent on prior
Regression Modelling	• Cause and effect relationship is required among variables • It is not suitable for larger data • Complicated procedure for the analysis • Poor quality data leads to unexpected results • It is not suitable for a larger number of parameters • Nonlinearity is not considered in this model • More suitable for numerical kind of data
Logistic Regression Modelling	• Overfitting is a major problem • Linearly separable data is difficult to find in real-world • Complex relationships are difficult to identify using this approach • Independent variables must be linearly related
Hypothesis Testing	• It is basically used to test the hypothesis, so not suitable for the proposed research • It focuses on the significance of the data, so estimation and confirmation are sometimes missed

The forecast of ordinal information is imperative for various applications. Machine learning procedures [19] play a vital part in expectations. Classification models [20] are appropriate for ordinal information as they contain a few classes. Different classification calculations are accessible that allow great come about in discrete datasets. For ordinal information, a particular calculation is required that addresses categorical information in a requested mould. Calculated relapse [21-23] could be a well-known calculation that bargains with categorical information viably. The other segment of the paper will bargain with the calculated relapse calculation in more detail.

5. Logistic Regression Model

One of the foremost broadly utilized machine learning methods, logistic regression, predicts the categorical dependent variable from a set of independent factors. In a categorical dependent variable, the output is predicted via logistic regression. Consequently, the result must be a discrete or categorical value. Sigmoid [24, 25], a scientific work, is utilized to outline the anticipated values to probabilities. The proposed research applies the threshold value idea in logistic regression, which establishes the likelihood of either 0 or 1. Examples include values that incline to 1 over the threshold value and to 0 below it. Logistic regression is best suited for binary classification. The mathematical equation of logistic regression is derived from a simple linear regression model [26, 27].

A simple linear regression equation is expressed as:

$$dv = c0 + c1iv1 + c2iv2 + \cdots + cnivn \tag{1}$$

As the result of logistic regression is between o and 1, divide dv by 1-dv. The result is 0 if dv=0, and it is infinite if dv=1. In order to get the result between - ∞ to +∞, take a log so it becomes:

$$\log\left(\frac{dv}{a-dv}\right) = c0 + c1iv1 + c2iv2 + \cdots + cnivn \tag{2}$$

The equation-2 is the logistic regression equation.

Basically, binary categorization is where it excels. It can, however, also be applied to ordinal data. The algorithm has several assumptions, such as:

- The nature of the dependent variable must be categorical.
- There should not be any multi-collinearity in the independent variable.

6. Proposed Ranking Learning Algorithm for Likert Scale (RLALS)

Assume that there are five outcomes, and the probability of them is represented as {p1(a),p2(b),p3(c),p4(d),p5(e)}. Probability p1(a) represents the minimum value while p5(e) represents the highest values. The probability minimum value is obtained by taking the logarithm of the division of p1(a) and summation of all other probabilities than p1(a). Similarly, way minimum value or next value is obtained by taking the logarithm of the division of p1(a) and p2(b) and the summation of all remaining probabilities.

Let's say the minimum value is represented as m1, and the maximum value is represented as m5.

$$m1 = \log((p1(a))/((p2(b) + p3(c) + p4(d) + p5(e)) \tag{3}$$

$$m1 \text{ or } m2 = \log((p1(a) + p2(b))/((p3(c) + p4(d) + p5(e)) \tag{4}$$

$$m1 \text{or } m2 \text{ or } m3 = \log((p1(a) + p2(b) + p3(c))/((p4(d) + p5(e)) \tag{5}$$

$$m1 \text{ or } m2 \text{ or } m3 \text{or } m4 = \log((p1(a) + p2(b) + p3(c) + p4(d))/(p5(e)) \tag{6}$$

A set of threshold values and linear functions of independent variables are used to estimate probability scores.

$$p(i) = p(l_{i-1} < \beta1a1j + \beta2a2j + \ldots\ldots + \beta lalj + Aj < l \ \leq_i) \tag{7}$$

Where β1, β2… βl are coefficient

l1,l2,….lj are threshold values

Aj is the assumption about the distribution

Value of l lies between -∞ to +∞

Based on Equation 7, the probability of the event is to be determined. Table 2 describes the detailed proposed algorithm.

7. Dataset, Experiments and Result Analysis

The data is collected from 339 students using five criteria related to curriculum, teaching-learning and research. The five criteria are depicted in Table 3. The five criteria are relevant to teaching-learning and research. Learning value in terms of skills, knowledge, and analytical abilities are considered in this criteria. The second criteria deal with the applicability or relevance of topics in real-life situations. The death of knowledge in terms of syllabus is included in the third criterion. The extent to which all topics are covered is considered as the fourth parameter. The last criterion is about the research project or dissertation relevance.

A total of 12 parameters, including criteria, are considered for the experimentation purpose. The parameters of the dataset are described in Table 4. The type of questions is closed-ended with different levels of categorical order such as Excellent, Very good, Good, Average, etc.

R programming language is used for devising an algorithm and performing experimentation. R is selected due to the number of characteristics such as open source, massive packages and libraries for data analysis and platform independence. A total of 130 samples of different characteristics are created for the experimentation. At the initial level, experimentation is carried out by considering all features. No feature engineering is carried out at the initial stage. Table 5 depicts the accuracy of the proposed model and logistic regression model before feature engineering. The overall accuracy of the RLALS algorithm is 68.63%, which is higher than the logistic regression algorithm by 7 percentages. The proposed model also performs well at individual Likert Scale levels defined in the dataset. Figure 1 describes the accuracy of the proposed model and logistic regression model by different Likert Scale levels given by the students for various criteria of teaching-learning, curriculum and research.

Table 2. Detailed proposed RLALS algorithm

Step 1 : Transform the prominent independent variables in numerical fields if they are in different types

Step 2 : Transform the dependent/ target variable into the ordered variable

as.ordered (Target Variable)

Step 3 : Build a Model based on the training dataset

Training Set = {Tr1, Tr2, Tr3,… Trn} // 70 to 80 % of data

Model 🡐 RLALS {Target Variable ~ prominent independent variables, Training Set}

Step 4 : Predict data based on testing data set

Testing Set = {Ts1Ts2, Ts3,… Tsm} // 20 to 30 % data

Predict 🡐 {Model, Testing Set}

Step 5 : Determine Accuracy of the model

Step 6 : Improvise the model based on appropriate feature engineering on the basis of p-values of independent variables.

$$Z \leftarrow (p^{\wedge} - p0)/\sqrt{[p0 - \frac{1-p0}{n}]}$$

Table 3. Criteria of curriculum, Teaching-Learning and research

Learning value (in terms of skills, concepts, knowledge, analytical abilities, or broadening perspectives
Applicability/relevance to real-life situations
Depth of the course content
Extent of coverage, of course
Relevance/learning value of project/report/ Research Dissertation

Table 4. Parameters of the dataset

Parameter	Description
Roll No	Unique Identification of Student Assigned by the Institution.
Name	Name of the Student.
Degree	Programme in which student is studying (i.e. BCA, B.Sc., IT, MCA, M.Sc., IT, etc.)
Gender	Gender of the Student (i.e. Male, Female, Transgender etc.)
Category	Category of the student (i.e. General, SC, OBC etc.)
Religion	Religion of the student.
Mother Tongue	Mother Tongue of the student.
Blood Group	Blood Group of the student.
Last Exam	The last Exam is given by the student.
Percentage	Percentage achieved in the last exam
Board/ University	Board or university of the last exam.
Scale of Criteria	Scale is given by the student in the criteria of Curriculum, Teaching-Learning and Research.

Table 5. Accuracy of proposed model and logistic regression before feature engineering

Model	Accuracy
Logistic Regression	61.69%
RLALS algorithm (Proposed Algorithm)	68.63%

Table 6. Accuracy of proposed model and logistic regression after feature engineering

Model	Accuracy
Logistic Regression	67.56%
RLALS algorithm (Proposed Algorithm)	89.24%

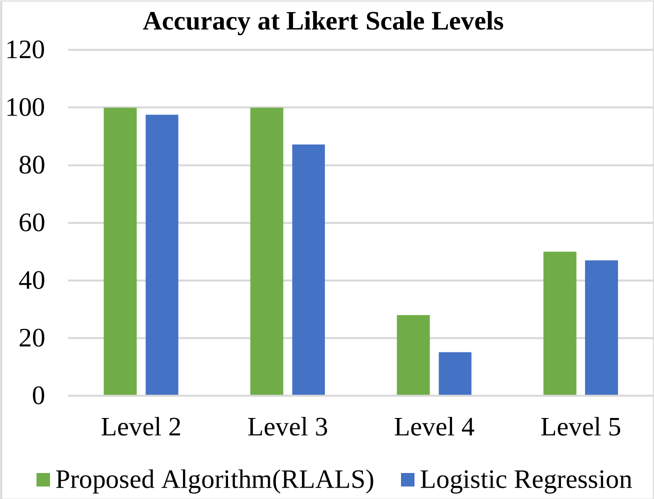

Fig. 1 Accuracy at likert scale levels of RLALS algorithm and logistic regression model

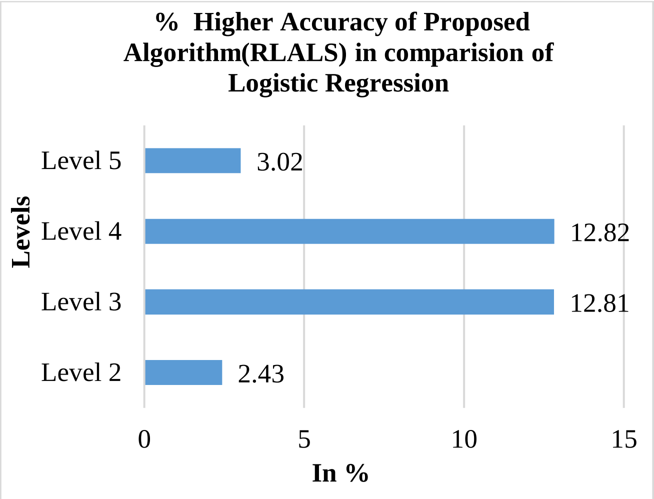

Fig. 2 Higher accuracy in % of RLALS algorithm in comparison to logistic regression

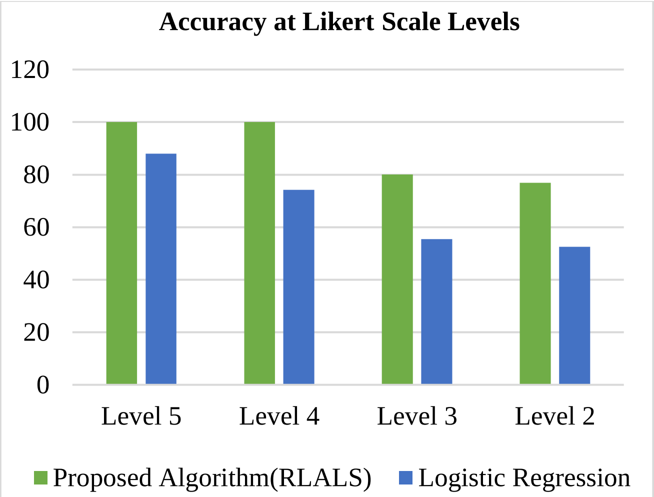

Fig. 3 Accuracy at likert scale levels of RLALS algorithm and logistic regression model after feature engineering

From Figure 1, it is determined that the proposed algorithm excels at every Likert scale level in comparison to the Logistic Regression algorithm. Figure 2 describes the percentage higher accuracy of the proposed algorithm at every level in comparison to the Logistic Regression algorithm. From Figure 2, it is determined that the proposed algorithm performs higher accuracy at every level. At level-3 and 4, the proposed algorithm produces more than 12% higher accuracy. The accuracy of level-2 and level-5 is also higher near about 3% of the proposed algorithm in comparison to logistic regression.

However, the overall accuracy of an algorithm is not satisfactory. Feature engineering plays an important role in order to select prominent features from the dataset. The proposed algorithm applies feature engineering on the basis of p-values and selects prominent features from the dataset. From the analysis of the p-value, it is determined that Gender, Category, Religion and Blood Group variables are irrelevant. Table 6 describes the accuracy of the proposed model and the well-known logistic regression model after feature engineering. The experiment was carried out with the same 130 samples. From the result, it is clear that the proposed algorithm performs excel after feature engineering. The accuracy of the proposed model is more than 21% higher than the logistic regression algorithm. The proposed algorithm suits ordinal data. From Figure 3, it is determined that the proposed algorithm excels at every Likert scale level in comparison to the Logistic Regression algorithm after feature engineering.

8. Conclusion

The Likert scale is vital for various domains to take opinions or feedback about something. This scale has categorical values but some intrinsic order, so able to be compared. Logistic regression is well well-known algorithm to predict class from several independent variables. Logistic regression suits binary classification but does not provide good results for ordinal kinds of data. In this paper, a new algorithm was proposed to predict Likert scale data. The new algorithm provided good prediction accuracy in comparison to Logistic Regression. The experiment was carried out with all independent variables and also with prominent independent variables. The proposed algorithm provided better results in both situations in comparison to the logistic regression algorithm. The experiment was done on education domain data. The algorithm is fitted with any kind of ordinal data of any domain.

Conflict of Interest

We hereby certify that, to the best of our knowledge, (1) the work which is reported on in said manuscript has not received financial support from any organization, and (2) neither We nor any first-degree relative has any special financial interest in the subject matter discussed in said manuscript.

References

[1] Arundhathi Thangeda, Bakisanani Baratiseng, and Thatoyamodimo Mompati, "Education for Sustainability: Quality Education is a Necessity in Modern Day. How Far Do the Educational Institutions Facilitate Quality Education?," *Journal of Education and Practice*, vol. 7, no. 2, pp. 9-17, 2016. [Google Scholar] [Publisher Link]

[2] Albert Victor Kelly, *The Curriculum: Theory and Practice*, 6th ed., SAGE Publications, pp. 1-336, 2009. [Google Scholar] [Publisher Link]

[3] Burcu Tezcan-Unal, Kalman Winston, and Anne Qualter, "Learning-Oriented Quality Assurance in Higher Education Institutions," *Quality in Higher Education*, vol. 24, no. 3, pp. 221-237, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[4] Renée Stalmeijer et al., "Strengthening Internal Quality Assurance Processes: Facilitating Student Evaluation Committees to Contribute," *Assessment & Evaluation in Higher Education*, vol. 41, no. 1, pp. 53-66, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[5] Alexander C. McCormick, Jillian Kinzie, and Robert M. Gonyea, *Student Engagement: Bridging Research and Practice to Improve the Quality of Undergraduate Education*, Higher Education: Handbook of Theory and Research, pp. 47-92, 2013. [CrossRef] [Google Scholar] [Publisher Link]

[6] I. Elaine Allen, and Christopher A. Seaman, "Likert Scales and Data Analyses," *Quality Progress*, vol. 40, no. 7, pp. 64-65, 2007. [Google Scholar] [Publisher Link]

[7] Harry N. Boone Jr, and Deborah A. Boone, "Analyzing Likert Data," *Journal of Extension*, vol. 50, no. 2, pp. 1-6, 2012. [CrossRef] [Google Scholar] [Publisher Link]

[8] Spencer E. Harpe, "How to Analyze Likert and Other Rating Scale Data," *Currents in Pharmacy Teaching and Learning*, vol. 7, no. 6, pp. 836-850, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[9] Paulo Soares, and Carlos Daniel Paulino, "Incomplete Categorical Data Analysis: A Bayesian Perspective," *Journal of Statistical Computation and Simulation*, vol. 69, no. 2, pp. 157-170, 2001. [CrossRef] [Google Scholar] [Publisher Link]

[10] Peter E Rossi, Zvi Gilula, and Greg M. Allenby, "Overcoming Scale Usage Heterogeneity: A Bayesian Hierarchical Approach," *Journal of the American Statistical Association*, vol. 96, no. 453, pp. 20-31, 2001. [CrossRef] [Google Scholar] [Publisher Link]

[11] Shaun R. Seaman, and Sylvia Richardson, "Equivalence of Prospective and Retrospective Models in the Bayesian Analysis of Case-Control Studies," *Biometrika*, vol. 91, no. 1, pp. 15-25, 2004. [CrossRef] [Google Scholar] [Publisher Link]

[12] Samiran Sinha, Bhramar Mukherjee, and Malay Ghosh, "Bayesian Semiparametric Modeling for Matched Case-Control Studies with Multiple Disease States," *Biometrics*, vol. 60, no. 1, pp. 41-49, 2004. [CrossRef] [Google Scholar] [Publisher Link]

[13] Tom Leonard and Hsu, J. S. J., "The Bayesian Analysis of Categorical Data: A Selective Review," In P. R. Freeman & A. F. M. Smith (Eds.), *Aspects of uncertainty: A tribute to D. V. Lindley*, pp. 283-310, 1994. [Google Scholar]

[14] Sander Greenland, "An Application of Logistic Models to the Analysis of Ordinal Responses," *Biometrical*, vol. 27, no. 2, pp. 189-197, 1985. [CrossRef] [Google Scholar] [Publisher Link]

[15] Werner Holtbrügge, and Martin Schumacher, "A Comparison of Regression Models for the Analysis of Ordered Categorical Data," *Journal of the Royal Statistical Society: Applied Statistics, Series C*, vol. 40, no. 2, pp. 249-259, 1991. [CrossRef] [Google Scholar] [Publisher Link]

[16] Cyrus R. Mehta, Nitin Patel, and Pralay Senchaudhuri, "Exact Stratified Linear Rank Tests for Ordered Categorical and Binary Data," *Journal of Computational and Graphical Statistics*, vol. 1, no. 1, pp. 21-40, 1992. [CrossRef] [Google Scholar] [Publisher Link]

[17] Cyrus R. Mehta, Nitin R. Patel, and Anastasios A. Tsiatis, "Exact Significance Testing to Establish Treatment Equivalence with Ordered Categorical Data," *Biometrics*, vol. 40, no. 3, pp. 819-825, 1984. [CrossRef] [Google Scholar] [Publisher Link]

[18] Mokhammad Ridwan Yudhanegara, "Carefully to Analyze the Data Type of Ordinal Scale, Why?," *Proceedings the 4th SEA-DR 2016*, pp. 215-221, 2016. [Google Scholar] [Publisher Link]

[19] Teng Xiuyi, and Gong Yuxia, "Research on Application of Machine Learning in Data Mining," *IOP Conference Series: Materials Science and Engineering*, vol. 392, no. 6, pp. 1-4, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[20] S.B. Kotsiantis, *Supervised Machine Learning: A Review of Classification Techniques*, Emerging Artificial Intelligence Applications in Computer Engineering, IOS Press, vol. 160, no. 1, pp. 3-24, 2007. [Google Scholar] [Publisher Link]

[21] Chang Zhenhai, and Liu Zuo, "Logistic Regression Model and its Application," *Journal of Yanbian University (Natural Science)*, vol. 38, no. 1, pp. 28-32, 2012. [Google Scholar] [Publisher Link]

[22] Michael P. LaValley, "Logistic Regression," *Circulation*, vol. 117, no. 18, pp. 2395-2399, 2008. [CrossRef] [Google Scholar] [Publisher Link]

[23] David G. Kleinbaum, Mitchel Klein, and Erica Rihl Pryor, *Logistic Regression: A Self-Learning Text*, 2nd ed., Springer, pp. 1-513, 2002. [Google Scholar] [Publisher Link]

[24] Xiaonan Zou et al., "Logistic Regression Model Optimization and Case Analysis," *2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT)*, Dalian, China, pp. 135-139, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[25] Jing Ren et al., "A Potential Field Model Using Generalized Sigmoid Functions," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 37, no. 2, pp. 477-484, 2007. [CrossRef] [Google Scholar] [Publisher Link]
[26] Keith A. Maril, "Advanced Statistics: Linear Regression, Part I: Simple Linear Regression," *Academic Emergency Medicine*, vol. 11, no. 1, pp. 87-93, 2004. [CrossRef] [Google Scholar] [Publisher Link]
[27] Mahmoud A. Mahmoud, J.P. Morgan, and William H. Woodall, "The Monitoring of Simple Linear Regression Profiles with Two Observations Per Sample," *Journal of Applied Statistics*, vol. 37, no. 8, pp. 1249-1263, 2010. [CrossRef] [Google Scholar] [Publisher Link]