Advances in Artificial Intelligence and Machine Learning; Research 3 (4) 1743-1757 Received 20-10-2023; Accepted 20-12-2023; Published 27-12-2023

Significance Statistical Test Analysis on Classification Models of Adolescent's Emotional Problems

Akhyt Tilyeubai akhit@mnums.edu.mn
Department of Physics & Medical Informatics, School of Biomedicine, Mongolian National University of Medical Sciences, Mongolia

Javzmaa Tsend javzmaa.ts@mnums.edu.mn
Department of Physics & Medical Informatics, School of Biomedicine, Mongolian National University of Medical Sciences, Mongolia

Bayarmaa Vaanchindorj
National Centre of Mental Health, Mongolia

Galbadrakh Chuluunbaatar
Department of Physics & Medical Informatics, School of Biomedicine, Mongolian National University of Medical Sciences, Mongolia

Baasandorj Chilkhaasuren
Department of Physics & Medical Informatics, School of Biomedicine, Mongolian National University of Medical Sciences, Mongolia

Ajnai Luvsan-Ish
Department of Physics & Medical Informatics, School of Biomedicine, Mongolian National University of Medical Sciences, Mongolia

Jargalbat Puntsagdash
Department of Physics & Medical Informatics, School of Biomedicine, Mongolian National University of Medical Sciences, Mongolia

Purevdolgor Luvsantseren
Department of Physics & Medical Informatics, School of Biomedicine, Mongolian National University of Medical Sciences, Mongolia

Bat-Enkh Oyunbileg
Associate Professor, Department of Information Technology, School of Information and Communication Technology

Corresponding Author: Javzmaa Tsend

Citation: Akhyt Tilyeubai, et al. Significance Statistical Test Analysis on Classification Models of Adolescent's Emotional Problems Advances in Artificial Intelligence and Machine Learning. 2023;3(4):100.

Abstract

In many countries, researches are being conducted to generate effective knowledge from big data using data mining methods. These methods have been tested on data such as air pollution, diabetes, cardiovascular, adolescent emotions, etc., creating valuable knowledge and contributing to the field of health sciences in Mongolia. We tested the decision tree algorithms on the data of children under five years of age and PM10 and PM2.5 fine particles for each month of 2019-2020, and the C50 method was highly effective in building and evaluating classification tree models.

Globally, one in seven people between the ages of 10 and 19 have a mental disorder, which is 13% of adolescents. The main causes of illness and disability in adolescents are depression, anxiety and behavioural disorders . On the report Mental Health System's in Mongolia of World Health Organization, for Mongolia, special attention needs to be given to develop professional competence and services in the area adolescent mental health and considered the need to expand mental health research and publish articles in indexed journals . Considered the need to expand mental health research and publish scientific articles in indexed journals.

Consequently, the SDQ were taken from students, class teachers and parents of the 6-12th grade of Govi-Altai Province to evaluate the student's emotions and created the student, parent-guardian, and teacher evaluation databases.

When divide the student evaluation database into ten using cross-validation and create the models by C50, Bayes, Ripper methods, evaluate by measures such as sensitivity, specificity, accuracy, t test, Bayes method model was showed good result.

Keywords: Data mining, classification, Model, Measure, Statistical tests.

1. INTRODUCTION

In recent years, rapid advances in internet technology and communications have continued to generate massive amounts of data. Data mining is transforming process from large amounts data in databases and data warehouses into effective information and knowledge. In Mongolia, classification and cluster analysis methods of data mining are being tested on data collected in many fields such as nature, weather and medicine, and knowledge is still being discovered.

We emphasize that in the study by Baochang Zhou and Zhanghong Qian [1], they developed core group medicine (CGM) in Mongolian Medicine (MM) for the treatment of pulmonary infectious diseases (PID) using association rules and hierarchical clustering analysis of data mining. Using hierarchical clustering algorithm in R programming, 43 high frequency drugs were clustered into six categories. It found that the core group medicines screened by association rule analysis were clustered into one group and it is illustrating more the importance of core group medicine and the accuracy of association rule analysis. Therefore, took Chebulae Fructus -Toosendan Fructus -Gardeniae Fructus as the core group medicine to further more explore its mechanism of action in the treatment of PID.

Consider that developed models that predict indoor PM2.5 concentrations for pregnant women who participated in a randomized trial of portable air cleaners for Ulaanbaatar in Weiran Yuchi's study. Using multiple linear regression and random forest regression, it was modelled indoor PM2.5 concentrations. The models have 447 independent of PM2.5 measurements and 87 potential predictor variables that obtained from outdoor monitoring data, questionnaires, home assessments and geographic data sets. MLR and RFR had similar performance in cross-validation ($R2 = 50.2\%$, $R2 = 48.9\%$ respectively). They developed blended models that combined the MLR and RFR, and the blended MLR model that included RFR predictions had the best performance (cross validation $R2 = 81.5\%$) [2].

The objective of Hwan-Jeu Y., Sarangerel D., Adilsaikhan M., Erdenebaatar's [3], research was which build an international pancreaticoduodenectomy database implemented with security and clinical rule supporting functions, which made the data-sharing easier and improve the accuracy of data. The decision tree model showed that elderly patients with pylorus-preserving have higher proportion of delayed gastric emptying. The decision tree model for the pancreatic fistula were demonstrated that cases with non-pancreaticogastrostomy reconstruction - body mass index (BMI) >29.65 or PG reconstruction - BMI>23.7 - non-classic PD have higher proportion of pancreatic fistula after PD.

We are developing many research projects with data mining classification methods. For example, when our research team build the models how to air pollutions such as PM2.5, PM10, NO effect to under five mortality using C50, CART algorithms and evaluate the sensitivity, specificity, the results of C50 method were good.

There are many surveys done using traditional mathematical method in Mongolian country. For example, in 2020, in the joint research project "Psychological Health of Adolescents" of Good Neighbors International Organization and Oyutolgoi LLC, mental disorders of adolescents were investigated. In 2016, data of 600 children aged 10-19 in Ulaanbaatar City, Darkhan-Uul, Arkhangai, Hovd, Dornod provinces were processed and analyzed using SPSS software, and current state of social development and maturity of adolescents were studied and needs were determined. Also, in 2013, according to Mongolian adolescent's emotional and behaviour survey of the National Center for Mental Health, 40% of them were with emotional and behavioral disorders. Because we studied the state of mental health of adolescents in Mongolia through data mining, it is the innovative aspect and the significance of this research.

So, In this study based adolescent's emotional data of Govi-Altai province, C50, bayes, RIPPER methods will establish models which predict student's emotional internal characteristics and evaluate by sensitivity, specificity, and select optimal model using statistical test.

2. MATERIAL AND METHODS

2.1 The Strengths and Difficulties Questionnaire (SDQ)

Adolescents undergo many biological and socio-psychological changes, and respond to any unpleasant event or information by self-protective emotional and behavioural changes. According to WHO, more than 13% of adolescents aged 10-19 have been diagnosed with mental illness. Anxiety

and depression account for about 40% of these diagnosed mental disorders [4, 5]. /The State of the World's Children 2021/

According to the study [6], it was found that 30.5% of adolescents have borderline problems, while 9% have disorders related to emotions and behavior. These adolescents exhibited various behavioral issues, including 73.8% frequently playing computer games, 50% engaging in lying, 48.8% involved in stealing, and 41.3% participating in fights. Additionally, they also experienced more emotional symptoms, with 47.5% displaying a lack of attention, 15% suffering from insomnia, and 15% experiencing headaches. It is also concerning that suicides among children and teenagers have increased in recent years. Therefore, if a child can receive timely mental healthcare during adolescence, it can prevent them from becoming developing mental disorders in adulthood.

There are many methods for detecting mental health problems in children, and in this study, we utilized Robert Goodman's SDQ method. This psychiatric questionnaire evaluates emotional symptoms, conduct problems, hyperactivity, peer relationships, and prosocial behavior issues through input from classroom teachers, parents, and students. The internal characteristics of a student are determined by evaluating emotional symptoms and peer relationship problems, while the external characteristics are specified by evaluating conduct problems and hyperactivity [7, 8].

In this study, data was collected from 6th-12th grade students, their parents, and class teachers in Gobi-Altai province primary schools using the SDQ questionnaire. This data was then modelled using the classification method to predict their internal characteristics.

2.2 Classification

As artificial intelligence technologies are widely introduced in our country hospitals and health professionals need to work with them, it is necessary to educate them about the basic concepts and methods of artificial intelligence and data mining. Since the main content of artificial intelligence and data mining has recently been included in the training of MNUMS, we believe that doctors and students should first study the basic methods of classification. They should then test these methods on data and finally make predictions using the most suitable method. Therefore, basic classification methods were applied to the data collected from adolescents through the questionnaires. This research is one of the many research projects based on data mining methods conducted in collaboration with health professionals. Therefore, this chapter briefly discusses classification and its methods.

Classification is a process of data analysis that extracts models describing important data classes. Data classification is a two-step process, consisting of a learning step and a classification step. For the first stage, a classification algorithm such as decision tree, bayes, rule-based method builds a classifier that analyses from a training set made up of database tuples and their associated class labels.

In the second step, the model is used for classification, and the accuracy of the classifier is estimated using a test set. The test set is independent of the training tuples and was not used to construct the classifier. The accuracy of a classifier on a given test set is determined by the percentage of test set tuples that are correctly classified by the classifier. To evaluate this, the class label of each test tuple is compared with the predicted class label from the learned classifier [9]. There are several

classification algorithms available, including C50, CART, CHAID for decision trees, bayes, and rule-based methods. In this study, we utilized C50, bayes, and rule-based methods.

The C5.0 algorithm is an extension of the C4.5 algorithm. The model works by splitting the sample based on the field with the maximum information gain. Each sub-sample defined by the first split is then split again, based on a different field, and the process repeats until the sub-samples cannot be split any further. After that, the lowest level splits are re-examined, and models that do not significantly contribute to the value are pruned [8, 10–13]. C5.0 uses the concept of entropy to measure purity and express homogeneous changes in the class attribute of the dataset. The minimum value 0 indicates complete homogeneity, while 1 indicates the maximum impurity for the sample.

Bayesian classification is based on Bayes' theorem and has been found to be comparable in performance with decision tree and selected neural network classifiers. It has also exhibited high accuracy and speed when applied to large databases [9, 14].

Rules are an effective way of representing information or knowledge. A rule-based classifier utilizes a set of IF-THEN rules for classification. The "IF" part (or left side) of a rule is referred to as the rule antecedent or precondition, while the "THEN" part (or right side) is the rule consequent. In the rule antecedent, the condition comprises one or more attribute tests that are logically ANDed. A rule's coverage is the proportion of tuples covered by the rule. To determine a rule's accuracy, we examine the tuples it covers and calculate the percentage of them that the rule can correctly classify [9].

IF-THEN rules can be extracted directly from the training data using a sequential covering algorithm. Rules are learned sequentially, where each rule for a given class will ideally cover many of the class tuples. There are many sequential covering algorithms such as AQ, CN2, and RIPPER Rules are learned one at a time. Each time a rule is learned, the tuples covered by the rule are removed, and the process repeats on the remaining tuples.

2.3 Model evaluation and selection

We have tried different methods to build more than one classifier and now will compare these by metrics of classifier performance such as accuracy, sensitivity, specificity [9]. That's why in this section, accuracy, sensitivity, specifity etc that evaluate classifier's performance were explained.

2.3.1 Metrics for Evaluating Classifier Performance

First, need to study the following terminology:

- True positives (TP) refer to the positive tuples that were correctly labelled by the classifier,
- True negatives (TN) are the negative tuples that were correctly labelled by the classifier,
- False positives (FP) are the negative tuples that were in correctly labelled as positive,
- False negatives (FN) are the positive tuples that were mis labelled as negative

These measures are summarized in TABLE 1's confusion matrix. The confusion matrix is a valuable tool for analysing the classifier's ability to recognize tuples from various classes.

Table 1: Confusion matrix displaying the total number of positive and negative cases.

	Predicted class			
		Yes	No	Total
Actual class	Yes	TP	FN	P
	No	FP	TN	N
	Total	P'	N'	P+N

For m classes (where m ≥ 2), a confusion matrix is a table of size at least m by m. An entry, CMi,j, in the first m rows and m columns indicates the number of tuples of class i that were classified as class j by the classifier. To achieve good accuracy, a classifier should ideally have most of the tuples represented along the diagonal of the confusion matrix, from entry CM1,1 to entry CMm,m, with the remaining entries being zero or close to zero. The table may include extra rows or columns to enhance accuracy, specificity, and sensitivity. For example, in the confusion matrix, P and N are shown. P represents the number of tuples that were labelled as positive (TP+FP), while N represents the number of tuples that were labelled as negative (TN+FN). The total number of tuples is TP+TN+FP +TN, or P+N. We will now learn about the evaluation measures, starting with accuracy.

The accuracy of a classifier on a given test set is the percentage of test set tuples that are correctly classified by the classifier.

$$Accuracy = \frac{TP + TN}{P + N} \tag{1}$$

The error rate or misclassification rate of a classifier is calculated as 1 minus the accuracy.

$$Error\ rate = \frac{FP + FN}{P + N} \tag{2}$$

When the distribution of the data set shows a significant majority of the negative class and a minority positive class, measures of sensitivity and specificity are considered. Sensitivity refers to the proportion of positive tuples that are correctly identified, while specificity refers to the proportion of negative tuples that are correctly identified.

$$Sensitivity = \frac{TP}{P} \qquad \text{and} \qquad Specificity = \frac{TN}{N} \tag{3}$$

2.3.2 Cross-Validation

Cross-validation are common techniques for assessing accuracy, based on randomly sampled partitions of the given data. In k-fold cross-validation, the initial data are randomly partitioned into k mutually exclusive subsets or "folds" D1, D2, …, Dk, each of approximately equal size. Training and testing are performed k times. In iteration i, partition Di is reserved as the test set, and the remaining partitions are collectively used to train the model. That is, in the first iteration, subsets D2, ..., Dk collectively serves as the training set to obtain a first model, which is tested on D1; the

second iteration is trained on subsets D1, D3, ..., Dk and tested on D2; and so on. The accuracy estimate is the overall number of correct classifications from the k iterations, divided by the total number of tuples in the initial data [9].

2.3.3 Statistical Tests of Significance

Many classifiers were studied in the previous section to understand how to create them, and in this section, we will focus on selecting the "best" one. We explored the use of statistical significance tests to evaluate if the variation in accuracy among the classifiers is a result of chance.

Cross-validation divides data into ten subsets and creates ten models using classification methods. To determine if there is any difference in the mean error rates of these models, we will use paired-samples t-test to determine whether there is a significant mean difference between two observations. Therefore, statements such as "Any observed mean will not vary by ± the standard errors 95% of the time for future samples" or "One model is better than the other by a margin of error of ± 4%" can be made [9].

In this study, the same test set can be used for both M1 and M2, because we performed pairwise comparison of the models during each round of 10-fold cross-validation. That is, for the itch round of 10-fold cross-validation, the same cross-validation partitioning is used to obtain an error rate $(err(M_1)_i, err(M_2)_i)$ for M1 and for M2. The error rates for M1 or M2 are then averaged to obtain a mean error rate for M1 or M2, denoted $\overline{err}(M_1)$ *or* $\overline{err}(M_2)$. The variance of the difference between these models is calculated. The t-test computes the t-statistic with k -1 degrees of freedom for k samples. In our study we have k = 10 since, here, the k samples are our error rates obtained from ten 10-fold cross-validations for each model. The t-statistic for pairwise comparison is computed as follows:

$$t = \frac{\overline{err}(M_1) - \overline{err}(M_2)}{\sqrt{var(M_1 - M_2)/k}} \tag{4}$$

$$var\,(M_1 - M_2) = \frac{1}{k}\sum_{i=1}^{k} \sqrt{err(M_1)_i - err(M_2)_i - \overline{err}(M_1) - \overline{err}(M_2)} \tag{5}$$

Let $err(M_1)_i$ be the error rate of model M_1on round i

Let $err(M_2)_i$ be the error rate of model M_2on round i

Let $\overline{err}(M_1)$ be mean error rate for M_1

Let $\overline{err}(M_2)$ be mean error rate for M_2

We want to determine if the difference between M1 and M2 is significantly different for 95% of the population, which corresponds to a significance level of 5% or 0.05. To do this, need to find the t-distribution value corresponding to k-1 degrees of freedom from the table. However, because the t-distribution is symmetric, typically only the upper percentage points of the distribution are shown. Therefore, the table value for t = sig/2 is looked up, which in this case is 0.025, where t is also referred to as a confidence limit. If t > z or t < -z, then our t value lies in the rejection region, within the distribution's tails. This means that the null hypothesis can be rejected, indicating that the means

of M1 and M2 are not the same, and we can conclude that there is a statistically significant difference between the two models. Otherwise, if the null hypothesis cannot be rejected, we conclude that any difference between M1 and M2 can be attributed to chance.

Our country's healthcare researchers and experts extensively utilize paired samples t-tests to compare the means of two samples using traditional mathematical methods. That's why this work is an innovative survey that applied the t-test to select classification models and introduced it to researchers, especially health researchers. In other words, our aim is to introduce data mining into health sector research in Mongolia by utilizing the well-known statistical criteria of researchers.

3. EXPERIMENT, RESULTS

In cooperation with NCMH, we have collected SDQ data from 3764 students, parents/guardians, and classroom teachers in Gobi-Altai province. This data has been used to create evaluation databases for students, guardians, and teachers. Each database includes attributes such as general information, demographics, emotional and behavioral problems, difficulties, and influencing factors. The sample size for this study is 3091.

Based on a cross-sectional study, with a confidence level of 98%, an error margin of 2%, a population proportion of 50%, and a population size of 10241, we can confidently say that 2549 or more samples are sufficient to represent the population. The population size is the adolescents of Gobi-Altai province.

We studied how borderline and abnormal states of emotions depend on indicators such as emotions and peer relationships in internal characteristics using the classification method on the student database.

The internal characteristics of students were evaluated based on the sum score of emotional factors (S3, S8, S13, S16, S24) and peer relationship factors (S6, S11, S14, S19, S23). Consequently, the InterCl class attribute was added to the database. The database consists of 3074 rows and a total of eleven attributes, with the class attribute being InterCl. The class label attribute has three distinct values: normal, borderline, and abnormal.

In TABLE 2, the InterCl is a class attribute that includes normal, borderline, and abnormal states of emotions. Other attributes represent symptoms related to emotional and peer relationships.

TABLE 3 shows the student evaluation database.

There are 1355 borderline, 1736 abnormal and 673 normal records in the evaluation database, and because the normal state is littler than the abnormal and the borderline records, because have selected abnormal, borderline records from the dataset. Because the borderline and the abnormal state in the dataset are respectively 44% and 56%, we assumed that class labels were uniformly distributed and possible to use the the dataset.

We divided the database into ten folds using cross-validation sampling. We built models M1, M2, and M3 using the C50, bayes, and RIPPER algorithms, respectively. We evaluated the models'

Table 2: Score of student's internal characteristic

No	Attributes	Not True	Somewhat True	Certainly True
Emotional problems scale				
1	Often complains of headaches… (I get a lot of headaches…) (S3)	0	1	2
2	Many worries… (I worry a lot) (S8)	0	1	2
3	Often unhappy, downhearted… (I am often unhappy….) (S13)	0	1	2
4	Nervous or clingy in new situations… (I am nervous in new situations…) (S16)	0	1	2
5	Many fears, easily scared (I have many fears…) (S24)	0	1	2
Peer problems scale				
1	Rather solitary, tends to play alone (I am usually on my own) (S6)	0	1	2
2	Has at least one good friend (I have one goof friend or more) (S11)	2	1	0
3	Generally liked by other children (Other people my age generally like me) (S14)	2	1	0
4	Picked on or bullied by other children… (Other children or young people pick on me) (S19)	0	1	2
5	Gets on better with adults than with other children (I get on better with adults than with people my age) (S23)	0	1	2
7	Internalising score	Normal - 1	Borderline - 2	Abnormal - 3
6	Class attribute (InterCl)	0-5	6	7-10

Table 3: Student evaluation database

S3	S8	S13	S16	S24	S6	S11	S14	S19	S23	InterCl
0	2	0	1	0	1	2	1	0	0	2
2	2	0	1	0	1	0	0	0	0	2
1	1	1	0	1	0	1	1	0	0	2
1	0	1	1	0	2	1	1	0	1	2
0	2	0	1	0	0	2	1	0	1	2
0	1	0	1	0	1	2	1	0	0	2
0	1	0	1	1	0	2	0	0	0	1
0	1	1	1	1	0	2	0	0	0	2
1	1	1	1	1	2	2	1	0	0	3
1	2	1	1	1	0	2	1	0	0	3
0	1	1	1	1	0	2	2	1	2	3

accuracy. Cross-validation is a resampling method where data is divided into approximately equal subsets for training and model creation. TABLE 4 displays the Confusion Matrix for each method.

Table 4: Confusion Matrix

			Predicted		
			2	3	Accuracy
C50 – M_1	Actually	2	36.7	6.6	**0.8620**
		3	7.2	49.5	
Naïve Bayes -M_2		2	39	6.5	**0.8869**
		3	4.8	49.6	
RIPPER-M_3		2	39	6.5	**0.8750**
		3	4.8	49.6	

Also, the models were evaluated using measures such as sensitivity and specificity (TABLE 5).

Table 5: Metrics for Evaluating Classifier Performance

Metrics	**Sensitivity**	**Specificity**	**Accuracy**
	2	**3**	
Students			
C50	0.836	0.882	0.86
Bayes	0.890	0.8920	**0.89**
RIPPER	0.888	0.865	0.88

We have created several classifiers and wanted to choose the "best" using statistical test, which is referred to as model selection. First, the error rates $err(M_1)_i$, $err(M_2)_i$ for each partition of the models were calculated, the mean error rates $\overline{err}(M_1)$, $\overline{err}(M_2)$ were determined. The errors rates calculated for each partition were assumed to be k samples, the variance $var(M_1 - M_2)$ is calculated.

A t-statistic with k-1 degrees of freedom was calculated for k samples. In our research work, k is 10. The significance t-test was used to determine the difference between M1, M2, was compared to the value of $z = -2.76$ from the distribution table corresponding to the 5%/2 significance level.

Because our t value lies in the rejection region, within the distribution's tails, this means that we can reject the null hypothesis. This indicates that the models are not the same, and we can conclude that there is a statistically significant difference between the models (TABLE 6).

The Bayesian model results are shown (TABLE 7).

4. DISCUSSION

Let's mention some of the research studies that have determined the mental health of Mongolian adolescents.

Ai Aoki, Ganchimeg T., Anudari T., and others conducted a study examining the socioeconomic and lifestyle factors associated with mental health problems in 4th-class students at public elementary

Table 6: Statistical tests of significance

No	C50$err(M_1)_i$	$Bayeserr(M_2)_i$	RIPPER$err(M_3)_i$
1	0.18	0.10	0.12
2	0.15	0.12	0.14
3	0.11	0.13	0.12
4	0.10	0.10	0.13
5	0.14	0.11	0.12
6	0.12	0.14	0.13
7	0.15	0.10	0.16
8	0.14	0.10	0.13
9	0.15	0.10	0.08
10	0.14	0.08	0.12
$\underline{err}$	0.14	0.11	0.13
	$\overline{err}(M_1) - \overline{err}(M_2)$		0.0010
	$\sqrt{var(M_1 - M_2)/k}$		0.01
	Z confidence limit		**2.76**
	T test		**3**

Table 7: Bayes model

State	0	1	2	0	1	2
Emotional				**Peer problems scale**		
S3				**s6**		
2	0.69	0.28	0.03	0.64	0.30	0.04
3	0.33	0.52	0.15	0.31	0.505	0.18
s8				**S11**		
2	0.18	0.62	0.19	0.06	0.16	0.77
3	0.03	0.46	0.5	0.05	0.22	0.71
s13				**s14**		
2	0.69	0.25	0.015	0.11	0.55	0.33
3	0.23	0.60	0.16	0.08	0.61	0.30
s16				**s19**		
2	0.32	0.55	0.12	0.80	0.17	0.01
3	0.09	0.51	0.38	0.46	0.43	0.10
s24				**s23**		
2	0.53	0.40	0.05	0.51	0.44	0.03
3	0.13	0.54	0.31	0.28	0.57	0.14

schools in one district in Ulaanbaatar. They used the Strengths and Difficulties Questionnaire (SDQ) and employed multivariate logistic regression analysis. The study included 2301 children, and it found that low maternal education, low household income, lack of physical activity, and excessive screen time were associated with internalizing problems.

Compared to our research, there are similarities in terms of the participants (children) and the use of the SDQ method. However, there are also several differences, such as the analysis of adolescent data using multivariate logistic regression analysis. Descriptive analysis was conducted using a one-sample t-test for the age variable and a chi-squared test for other variables. The explanatory variables in the logistic regression analysis included gender, living area, family structure, maternal education, household income, sleep, physical activity, and screen time. Additionally, our study was conducted in the Gobi-Altai province, whereas this study focused on the 4th grade students in one district of Ulaanbaatar city [15].

Ai Aoki and Ganchimeg Togoobaatar aimed to investigate the association between children with mental health problems and their maternal quality of life (QOL), as well as other related factors in Mongolia. They conducted the study using the methods described in a previous study [16].

Amarjargal Dagvadorj a, Daniel J. Corsi a, Narantuya Sumya et al assessed the protective and risk factors for mental health problems in six-year-old children using logistic regression with adjustment for potential confounders [17].

Psychological Distress among Adolescents in four low- and middle-income countries (LAMICs) in Asia was explored and examined the relationship between this distress and the factors in individual, family, and school. Data analysing was applied descriptive statistics, Chi-squared testing, and logistic regression [18].

The suicidal plans and attempts based 5,393 adolescents data in Mongolia were specified to relate many factors such as urban and rural location and personal characteristics by logistic regression analyses [19].

Compared to the previous similar studies, our research have the advantage of specifing the mental health state of Mongolian adolescents using data mining methods. The previous survey differ in explored based on traditional statistic analysis such as descriptive statistics, statistic test, regress analysis. WHO-AIMS Report on Mental Health System in Mongolia was noted there is no mental healthcare for adolescents and in the future, special attention should be paid to the development of professional competence, research to provide mental health services for adolescents. This is consistent with the results of our research.

We have also compared our research to similar international studies.

In the study of Mining a Small Medical Data Set by Integrating the Decision Tree and t-test [20], 212 consecutive patients treated at the outpatient gynaecological department of Chang Gung Memorial Hospital, Taipei, Taiwan from July 1994 to July 2008 and all patients had undergone previous surgical treatment for ovarian endometriomas and were being seen because of a recurrence. Recurrence was defined as when one or more persistent pelvic cysts greater than 3.0 cm were detected in two consecutive ultrasonographic examinations. Transvaginal ultrasound-guided cyst aspiration and ethanol injection were done on an outpatient basis. According to the rule "Cyst_size ≥ 4.25 ==> Recovery = 0" all patients was divided into two groups by the cut point of the Cyst_size attribute. Group 1 contains patients whose cyst size is less than 4.25, while cyst size of patients in group 2 are not less than 4.25. All patients were divided into two groups: group 1 (Cyst_size <4.25) and group 2 (Cyst_size ≥ 4.25). It concluded the recovery rate of group 1 (43/59, 72.88%) is significantly (p = 1.47E-08) greater than that of group 2 (49/153, 32%) and in other words, the

recovery rate (or recurrence rate) is affected by the cyst size regardless of the treatment type. To deal with more complex decision trees, wanted to integrate other data mining the approaches to analyze postoperative status of ovarian endometriosis patients in the future. In this study, the p value was used to determine the difference between nodes in a single decision tree model constructed by the CART algorithm. Our study has the advantage of using t test to select the best among the classification models.

Pablo M., Emad M [21], studied on investigating different classification models developed to predict the Bridge Condition Index (BCI) in the province of Ontario, Canada, based on the publicly available historical data for 2,802 bridges over a period of more than 10 years. Predictive models used in this study include k-nearest neighbours (k-NN), decision trees (DTs), linear regression (LR), artificial neural networks (ANN), and deep learning neural networks (DLN). Models were evaluated by several performance metrics such as the root mean square error (RMSE), the mean absolute error (MAE), the mean relative error (MRE). Checking the variation of the RMSE of the proposed models, the error is reduced in a 78.38%, 80.66%, 80.15%, 87.10%, and79.28% for the DLN, LR, ANN, 4-NN, and DT models respec-tively. Similarly, the absolute and relative error of the models decreased by 89.18%, 88.51%, 92.24%, 92.30%, and 96.68%, re-spectively. The linear regression model was discarded as it has the highest errors. Performance-wise, 4-NNand DT seem to yield slightly more accurate predictions than the DLN and ANN models proposed, however, the differences are so small that a statistically-backed decision cannot be taken easily without further testing. As such, the pairwise t-test was applied to check the similarity between models from a statistical point of view and it was determined by looking at the p-value. The hypothesis would be validated with p-values that are close to zero for the paired predictive models, while values close to 1 would invalidate the hypothesis. A standard significance level of 5% is given usually to validate the proposed hypothesis of a t-test (p-value ≤ 0.05). In this case, the t-test results in validate the hypothesis for all the proposed mod-els. Therefore, from a statistical viewpoint, the performance of the proposed models is very comparable. Based on the results obtained with the tuned dataset, the performance evaluation tests on the different models, the authors recommended the use of the DT model for BCI prediction in the proposed case study. The study has the advantage which used k-nearest neighbors (k-NN), artificial neural networks (ANN), and deep learning neural networks (DLN) et al of advanced classification method. But it is same to using a paired sampling method to compare the classifiers. But it is same to using a paired sampling method to compare the classifiers.

For determining the difference between Neural Network and C45 algorithm, Thomas [22], reviewed statistical tests such as a test paired-differences t test, McNemar's, t test based on 10-fold cross-validation, a test for the difference of the two proportions, 5×2 cv. The 5x2cv test was the most powerful among those statistical tests. McNemars test does not assess the effect of varying the training sets, it still performed very well. The 5x2cv test will fall in cases where the error rates measured 2-fold cross validation replication vary wildly. The 10-fold cross validated t test has high Type I error. However, it also has high power, and hence, it can be recommended in those case where Type II error.

Researchers in the health sector of our country primarily utilize regression methods and compare models through statistical tests. The novelty of our research lies in testing significance tests on models constructed using the classification method. In our study, we employed the t-test for statistical significance to ascertain the actual difference in the mean error rate of the models.

5. CONCLUSION

- The accuracies of the models using cross-validation were 10% higher than those of the two splits with a ratio of 70:30, ranging from 0.86 to 0.89. This shows that cross-validation works better when there is more data
- According to the results of the statistical test, the model with the least error can be selected because the models are statistically different from each other. Therefore, the model accuracy calculated by the Bayesian method is 89%, which is 2% higher than the other models. Additionally, the error is small, so the results were considered reliable.
- According to the students evaluation who participated in our study, the their emotional borderline is likely when they sometimes worry about many things, when they are unhappy and depressed, when they are nervous and anxious, when they are afraid of many things.
- The borderline in peer relationships is likely to occur when they play alone, and the abnormal state is likely expect when having one close friend, other children sometimes do not like it, having peers.
- In the future, advanced classification methods, such as Bayesian belief networks and backpropagation, will be tested on health data. It is believing that these methods will provide more accurate results.

References

[1] Zhou B, Qian Z, Li Q, Gao Y, Li M, et al. Assessment of Pulmonary Infectious Disease Treatment With Mongolian Medicine Formulae Based on Data Mining, Network Pharmacology and Molecular Docking. Chin Herb Med. 2022;14:432-448.

[2] Yuchi W, Gombojav E, Boldbaatar B, Galsuren J, Enkhmaa S, et al. Evaluation of Random Forest Regression and Multiple Linear Regression for Predicting Indoor Fine Particulate Matter Concentrations in a Highly Polluted City. Environ Pollut. 2019;245:746-753.

[3] Yu HJ, Lai HS, Chen KH, Chou HC, Wu JM, et al. A Sharable Cloud-Based Pancreaticoduodenectomy Collaborative Database for Physicians: Emphasis on Security and Clinical Rule Supporting. Comput Methods Programs Biomed. 2013;111:488-497.

[4] https://www.who.int//news-room/fact-sheets/detail/adolescent-mental-health/?gclid=CjwKCAjws9ipBhB1EiwAccEi1IEoXV9oiaaE_MjLFHFLixczOnGjdMGAckEW4wYKbGksMwiN9SKTTBoC4xMQAvD_BwE

[5] https://cdn.who.int/media/docs/default-source/mental-health/who-aims-country-reports/mongolia_who_aims_report.pdf?sfvrsn=6421b370_3&download=true

[6] Bayarmaa V, Tuya N, Batzorig B, Ye G, Altanzul N, et al. Using the Strengths and Difficulties Questionnaire (Sdq) to Screen for Child Mental Health Status in Mongolia. J Ment Disord Treat. 2017;3:2.

[7] Goodman R. The Strengths and Difficulties Questionnaire (SDQ). In VandeCreek L, Jackson TL, editors, Innovations in clinical practice: Focus on children & adolescents. Sarasota, FL: Professional Resource Press/Professional Resource Exchange, Inc. 2003:109 - 111.

[8] Nternalizing Problems, Externalizing Problems, and Prosocial Behavior–-Three Dimensions of the Strengths And Difficulties Questionnaire (SDQ): A study among South African adolescents. Scand J Psychol. 2022;63:415-25.

[9] Han J, Kamber M, Pei J. Data Mining Concepts and Techniques. Third Edition. University of Illinois at Urbana-Champaign Micheline Kamber Jian Pei Simon Fraser University. 2012.

[10] Balamurugan M, Kannan S. Performance Analysis of Cart and C5. 0 Using Sampling Techniques. In: IEEE International Conference on Advances in Computer Applications (ICACA). IEEE Publications. 2016:72-75

[11] Palmov SV, Miftakhova AA. Comparison of classification algorithms C4.5 and C5.0. Infokommunikacionnye tehnologii. 2015;13:467-471.

[12] https://rstudio-pubs-static.s3.amazonaws.com/404024_4e62fe44761a4bc690918f93ac2a2aed.html

[13] Benediktus N, Oetama RS. The Decision Tree C5. 0 Classification Algorithm for Predicting Student Academic Performance. Ultimatics: Jurnal Teknik Informatika. 2020;12:14-19.

[14] https://docplayer.net/30552820-Naive-bayesian-classifier.html

[15] Aoki A, Togoobaatar G, Tseveenjav A, Nyam N, Zuunnast K, et al. Socioeconomic and Lifestyle Factors Associated With Mental Health Problems Among Mongolian Elementary School Children. Soc Psychiatry Psychiatr Epidemiol. 2022;57:791-803.

[16] Aoki A, Togoobaatar G, Tseveenjav A, Nyam N, Zuunnast K, et al. Quality of Life of Mothers of Children and Adolescents With Mental Health Problems in Mongolia: Associations With the Severity of Children's Mental Health Problems and Family Structure. Glob Ment Health (Camb). 2022;9:298-305.

[17] Dagvadorj A, Corsi DJ, Sumya N, Muldoon K, Wen SW, et al. Prevalence and Determinants of Mental Health Problems Among Children in Mongolia: A Population-Based Birth Cohort. Glob Epidemiol. 2019;1:100011.

[18] Lee H, Lee EY, Greene B, Shin YJ. Psychological Distress Among Adolescents In Laos, Mongolia, Nepal, and Sri Lanka. Asian Nurs Res. 2019;13:147-153.

[19] Davaasambuu S, Batbaatar S, Witte S, Hamid P, Oquendo MA, et al. Suicidal Plans and Attempts Among Adolescents in Mongolia. Crisis. 2017;38:330-343.

[20] Chang MY, Shih C, Chiang D, Chen C. Mining a Small Medical Data Set by Integrating the Decision Tree and T-Test. J Softw. 2011;6:2515-2520.

[21] Martinez P, Mohamed E, Mohsen O, Mohamed Y. Comparative Study of Data Mining Models for Prediction of Bridge Future Conditions. J Perform Constr Facil. 2020;34:04019108.

[22] Dietterich TG. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. Neural Comput. 1998;10:1895-1923.