BULGARIAN ACADEMY OF SCIENCES

CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume **22**, No 2
Sofia • 2022 Print ISSN: 1311-9702; Online ISSN: 1314-4081
DOI: 10.2478/cait-2022-0023

A Rule-Generation Model for Class Imbalances to Detect Student Entrepreneurship Based on the Theory of Planned Behavior

Nova Rijati[1], *Diana Purwitasar*[2], *Surya Sumpeno*[3], *Mauridhi Hery Purnomo*[3]

[1]*Department of Informatic Engineerings, Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang, Indonesia*
[2]*Department of Informatics, Faculty of Intelligent Electrical and Informatics Technology, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia*
[3]*Department of Computer Engineering, Faculty of Intelligent Electrical and Informatics Technology, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia*
E-mails: nova.rijati@dsn.dinus.ac.id diana.purwitasari@gmail.com surya@ee.its.ac.id hery@ee.its.ac.id

Abstract: *The ability to identify the entrepreneurial potential of students enables higher education institutions to contribute to the economic and social development of a country. Current research trends regarding the detection of student entrepreneurial potential have the greatest challenge in the unequal ratio of datasets. This study proposes a rule-generation model in an imbalanced situation to classify student entrepreneurship based on the Theory of Planned Behavior* (*TPB*)*. The result is a ruleset that is used for the early detection of student entrepreneurial potential. The proposed method consists of three main stages, namely preprocessing data to classify data based on TPB variables, generating a dataset by clustering and selecting attributes by sampling to balance the data, and finally generating a ruleset. Furthermore, the results of the detecting ruleset have been evaluated with actual data from the student tracer study as ground truth. The evaluation results show high accuracy so that the ruleset can be applied to the higher education environment in the future.*

Keywords: *Rule generating model, student entrepreneurial potential detection, imbalanced data, theory of planned behavior.*

1. Introduction

The entrepreneurial potential developed early can accelerate the economic growth of a Nation. Education also needs to be designed to handle this, especially higher education. To design this policy in the educational environment, it is necessary to know the main factors influencing student entrepreneurship interest. One approach to understanding these factors is the Theory of Planned Behavior (TPB) [1], in which entrepreneurial activity [2-5] is preceded by intention. The influencing variables according to this theory are Attitude (Att), Subjective Norm (SN), and Perceived

Behavioral Control (PBC). These variables in entrepreneurship education policies positively and significantly affect students' entrepreneurial interest [6-8].

Student entrepreneurial potential presents specific data patterns that indicate students' potential level and interest based on TPB variables generated from academic databases [9]. However, the number of students who consider entrepreneurial activities as a career choice is generally minor compared to those who do not. Suppose that situation is implemented as a classification problem using student's academic records as features along with binary labels of entrepreneurs and non-entrepreneurs. In that case, the classifier will have an imbalanced problem. This is commonly the case in education data mining research [10, 11] and can produce low accuracy in the detection of minority populations [12]. Methods for dealing with class imbalances are divided into three categories [13]. The first category is at the data level, and attempts to balance data distribution with oversampling and undersampling, such as Synthetic Minority Oversampling TEchnique (SMOTE) [14]. The second approach is at the algorithm level and employs the calculation of minority classes to understand data representation. That is, combinations of SMOTE techniques are performed on a filtered dataset with TPB variables to improve the accu racy performance of the classifier with binary labels related to student entrepreneurial potential.

A decision tree-based classifier could generate rules using data features. In the problem of classifying student entrepreneurial potential with TPB variables, those generated rules may assist in the design of entrepreneurship education policy. This study focuses on generating those rules based on TPB variables in imbalanced situations. This study also investigates the quality of generated rules from those algorithms on student entrepreneurial behavior data sampled with various sampling techniques to address the imbalanced data issue. It is expected that better-generated rules would make more explicit policy rules for designing strategies that encourage young entrepreneurs who graduated from university.

2. Related works

The theory of planned behavior [1] characterizes that the explanation of behaviors such as entrepreneurship is influenced by certain intended factors of Attitude (Att), Subjective Norm (SN), and Perceived Behavioral Control (PBC). Since Att, SN, and PBC have a positive and significant effect on entrepreneurial interests [4, 5], TPB contributes to understanding student intentions in forming entrepreneurial behavior.

The Att factor contributes a positive or negative response to a given assessment, like having cognitive and affective dimensions with different effects on entrepreneurial interest [3]. The SN factor refers to external factors that influence an individual to do or not perform a behavior. The SN value would be high when there are normative beliefs and motivation to fulfil the expectations of related people, such as pressure from peers, society, institutions, or other external environments. The SN factor has a more significant effect than Att and PBC on entrepreneurial interest. The PBC factor is an individual's perception in realizing a particular behavior [15]. This

factor could increase the strength of an individual's intention to perform a behavior or serve as an obstacle.

TPB has been integrated with fuzzy logic to measure entrepreneurial interest in the Global Entrepreneurship Monitor (GEM) [16], a survey conducted on adult populations and experts to uncover the entrepreneurial potential of students based on the GEM entrepreneurship process model. In our previous studies, TPB integration with a fuzzy-based approach of the Multi-Attribute Decision Making (MADM) technique in education datasets have been used to map student entrepreneurial potential based on their academic activities [9]. Our studies have displayed that K-Means clustering and Fuzzy MADM produce consistent data interpretations with the GEM dataset. Our preliminary studies have generated the TPB variables of Att, SN, and PBC from an academic database (Fig. 1) to demonstrate those interrelated beliefs. The clustering method could discover data structures to better understand preference relations, such that its usage and decision tree in educational data mining reveales hidden patterns of the student performance [17]. This method also could be combined with regression to group student behaviors of expert, good, regular, bad, and criticism [18]. We have used κ-Means clustering to observe the behavior of student entrepreneurial potential patterns from an academic database using Simple Multi-Attribute Rating Technique (SMART) preprocessing [19]. Then, preliminary experiments with a defined TPB model have been performed to understand the relation between entrepreneurial potential and student attributes [9].

As mentioned with a case in the Introduction section, the need to find student entrepreneurial potential leads to a problem of imbalanced data. Imbalanced data sampling can be performed by balancing the original data with the combination of oversampling and or under-sampling. Oversampling duplicates data from minority classes such as SMOTE (Synthetic Minority Oversampling Technique) that improves the classifier accuracy for minority classes [20], while under-sampling removes some data from majority ones. By adding the Nominal Continous (NC) feature to the SMOTE technique, the sampling method is called SMOTE-NC.

Some studies have compared several oversampling and under-sampling methods like the Cluster Centroid (CC) [21]. Random UnderSampling (RUS) method randomly and uniformly balances class distributions. It could cause information loss as the majority of the class instances are near each other, while NearMiss (NM) considers near-neighbor methods to overcome the potential information loss [22]. Other well-known sampling methods are the Edited Nearest-Neighbor rule (ENN), Repeated Edited Nearest Neighbor rule (RENN), or an ENN-based method of AllKNN. ENN reduces pre-classified samples by deleting closest neighbors to improve the classification accuracy of minority instances [23]. RENN repeatedly applies the ENN method until all remaining examples have most neighbors with the same class, and AllKNN deletes samples close to the minority class to separate existing classes [24]. There are also Tomek Link (Tomek or TL) that determines boundary classes, Condensed Nearest Neighbor (CNN) that finds paired points in linear boundaries [25], One-Sided Selection (OSS) that incorporates TL and CNN [26]. In addition is the Instance Hardness Threshold (IHT) that removes complex samples to avoid possible misclassification [27]. Previous studies have concluded

that SMOTE+Tomek and SMOTE+ENN could deliver better results for datasets with few positive class examples [22]. Since a small dataset has been applied to entrepreneurial students, we have executed sampling methods in the data preprocessing stage, before finding rules with a classification approach in this study.

3. Proposed method

This research proposes a TPB-based rule-generation method to detect student entrepreneurship with unequal data problems that need data sampling to obtain a more balanced dataset. The model proposed in Fig. 1 has three main stages, namely: dataset preprocessing, sampling dataset, and ruleset generation. The first stage performs data filtering based on TPB, the second stage makes clustering to produce the best sampling dataset, and the last stage obtains a detection model.

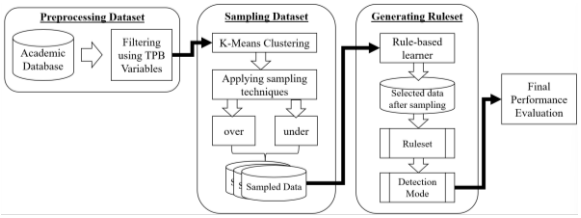

Fig. 1. The proposed framework of rules generator for detecting student entrepreneurship potential

Based on Fig. 1, conceptually, it can be explained that the proposed framework is based on a development model of a machine learning system trained to predict students' entrepreneurial status using a variable-based dataset (TPB) acquired from a higher education database. The knowledge used to predict is generated from a dataset of student entrepreneurship potential that has been validated using a sampling technique combined with attribute selection techniques. The entrepreneurial potential of students is identified based on the results of ranking data patterns using fuzzy hybrid computing that combines fuzzy logic with multi-criteria decision-making techniques. Furthermore, the data patterns are transformed into a knowledge base consisting of rules using reasoning adapted from the classification algorithm. The experiment generated two rulesets of prediction rules based on two dataset variants. All rulesets align with the concept of TPB behavior, namely forming a decision tree rooted in perceived behavioral control variables that can directly determine student entrepreneurial behavior. In detail, each main step of the method is described in Sections 3.1-3.3.

3.1. Pre-processing dataset

Our studies have utilized a dataset obtained from an Indonesian university's higher education database containing students, lecturers, and study programs called the academic database. After eliminating some inconsistent and incomplete data in the dataset preprocessing (see Fig. 1), there were 14 selected attributes to represent the student characteristics based on TPB variables that influence entrepreneurial interest. The dataset distribution related to student entrepreneurial potential has been as follows: the entrepreneurial students (4%) compared to those not entrepreneurial

(96%). There have been varying levels of entrepreneurial potential from those students with some scenarios using feature combinations of TPB variables: Medium 34%, High 55%, and Very High 11% [9].

The dataset in our experiment is taken from a university academic database focusing on undergraduate students from one cycle of four-academic years. Thus, we have observed entrepreneur potential from their first year until graduation, with some students taking more years before graduating (2009-2015, about 3000 students). Since we aimed to find out the possibility of entrepreneur potential, the dataset only contained 336 who joined the student entrepreneurial program offered by the government. The joining status is represented with a College Student Entrepreneur Program (CSEP) attribute. Since not all students who have joined that program would eventually become entrepreneurs, that is the motivation of our current works.

3.2. Sampling dataset

Our preliminary works with an academic applied database (SMART) preprocessing [19], has studied attribute selections [28], and then it has proposed a TPB-based model for student entrepreneurial potential [9]. Although the model could recognize student profiles with high potential to become entrepreneurs, the student population distribution of entrepreneurs does not show unequal proportions or imbalanced data. Thus, the model in Fig. 1 describes sampling steps for handling imbalanced data before generating rules. Data balancing has been performed on a ratio of 2:1 from the selected dataset after clustering. The sampling methods were oversampling with the SMOTE-NC technique and under sampling with 10 various techniques such as CC, CNN, ENN, RENN, AllKNN, IHT, NM, OSS, RUS, and TL. The processes of clustering and balancing data are performed to improve the preprocessing stage before generating rules for the detection model.

3.3. Generating rules and model evaluation

The extracted TPB variables related to the student environments are used to create rules as a detection model from decision tree methods called rule-based learners. There are two techniques for generating rules from previous studies: rule-based algorithms (e.g., Projective Adaptive Resonance Theory called as PART, 1R called as OneR, Repeated Incremental Pruning to Produce Error Reduction called as RIPPER) and decision trees (e.g., ADTree, J48 from C4.5, Reduced Error Pruning Tree called as REPTree) [29]. Those algorithms perform different functions, namely, the OneR algorithm only produces one rule, J48 is derived from a classic C4.5 algorithm, REPtree is a fast decision tree learning algorithm based on information gain or the variance reduction, the PART algorithm creates recursive rules for all data instances, and RIPPER uses a data generalization process to create the rules. Our process for generating rules using those algorithms of rule-based learners aims to formulate valid and effective rules for student entrepreneurial detections.

Decision tree classifiers include TPB variables, and the outputs indicate rules describing the relationship of academic behavior attributes that affect a student's interest in entrepreneurship. The performance evaluation of a classification rule R to detect student entrepreneurial potential utilizes coverage and accuracy indicators.

Coverage is the percentage of tuple coverage that conforms to a rule (Equation (1)). Accuracy is the percentage of covered and correctly classified (Equation (2)). From both equations of performance indicators, R is rule, n_{covers}: number of covered tuples, $|D|$: number of all tuples in a dataset, ncorrect: number of correctly classified tuples.

The proposed model for generating detection rulesets is evaluated using a confusion matrix to measure the model's performance regarding Precision, Recall, F-Measure, and G-Mean [30]. F-Measure is used to measure minority rates in imbalanced classes, and the G-mean index is used to measure overall classification performance.

(1) $$\text{coverage}(R) = \frac{n_{\text{covers}}}{|D|},$$

(2) $$\text{accuracy}(R) = \frac{n_{\text{correct}}}{n_{\text{covers}}}.$$

4. Results and discussion

4.1. Preparing the experimental dataset

The dataset used consists of 336 students with 14 attributes grouped into TPB variables of Att, SN, and PBC [9].

Table 1. Student entrepreneurial potential criteria

TPB Variable	Code	Criteria	Subcriteria		Weight
			Abbreviation	Description	
Att	A1	Scholarship status	Sch	Scholarship	1.00
			NonSch	Non-scholarship	0.50
	A2	Activist status	Act	Activists	1.00
			NonAct	Non-activists	0.50
	A3	Type of entry	NR	Non-regular	1.00
			Average (AVG)	AVG > 7.0	0.80
			RT	Regular Test	0.60
			T	Transfer	0.40
			M	Moving	0.20
SN	A4	Parents occupation	Entre	Entrepreneur	1.00
			NonEntre	Non-entrepreneur	0.50
	A5	Parents income	P1	> 10 million	1.00
			P2	7 – 10 million	0.80
			P3	5 – 7 million	0.60
			P4	3 – 5 million	0.40
			P5	< 3 million	0.20
	A6	Grade Point Average (GPA)	C	Cumlaude	1.00
			HS	Highly Satisfactory	0.75
			S	Satisfying	0.50
			G	Good	0.25
	Course Score				
	A7	Indonesian language	A B C D E	There are five grades of course scores	1.00 0.80 0.60 0.40 0.20
	A8	English			
	A9	Research method			
	A10	Professional ethics			
	A11	Counseling			
	A12	Entrepreneurship			
PBC	A13	Business incubator status	member	member	1.00
			non member	non member	0.50
	A14	College Student Entrepreneur Program (CSEP) status	E	Excellent	1.00
			F	Funded	0.67
			P	Proposing	0.33

In Table 1: A1-A3 describe attitudes that affect entrepreneurial activity (Att); A4-A12 describe learning processes in higher education (SN); A13-A14 describe student activities related to entrepreneurial events programmed and facilitated by the university (PBC). The SN variables relate to our research objective, which investigates the influence of educational processes in nurturing the entrepreneurial interest of students. The students in the dataset were rated with varying levels of entrepreneurial potential like medium, high, and very high. However, after a preprocessing phase that handled missing values, the composition of 277 students was determined to be 250 students with non-entrepreneurs status (90%) and 27 students with entrepreneur status (10%).

The information related to the students' parents, such as their occupation and income, describes their financial ability, which correlates to the TPB variable of SN. Therefore, even though A4 (parent occupation) and A5 (parent income) variables are not directly related to higher education's learning processes, they could become background motivation that encourages students to become entrepreneurs.

4.2. Examining student patterns with statuses of "entrepreneur" and "non-entrepreneur"

Previous studies [9] using k-Means clustering produced four clusters tested in seven scenarios as combinations of TPB variables: Att+SN+PBC, Att, SN, PBC, Att+PBC SN+PBC, and Att+SN. The experiments determined the appropriate combinations of TPB variables in forming clusters. The performances of clustering results based on Silhouette scores for those seven scenarios were 0.14, 0.69, 0.20, 0.95, 0.40, 0.17, and 0.16, respectively. It showed that Att attributes and PBC attribute alone give the best clustering results, while clustering with all attributes produces the lowest Silhouette score. The next clustering process had been performed on Att+PBC attributes that gave the next best Silhouette score (0.40) with $K = 2, \ldots, 6$ to observe the data consistency and the distribution of data members.

4.3. Forming a balanced dataset with over-sampling and under-sampling techniques

The following process balanced the data with sampling techniques to achieve the desired balance target: a data ratio of the entrepreneur to non-entrepreneurs as 1:2. From Fig. 2, clusters C1 and C3 have the most entrepreneurial student members than C2 and C4. Thus, synthetic data is generated with targets C1 and C3 using oversampling methods. The data generation with the SMOTE-NC oversampling technique produced 30 synthetic data samples for C1 and 28 – for C3 to make more balanced clusters of C1 and C3 having 40 data samples. After oversampling, the synthetic data (58) were merged with the initial data (227), and the current dataset had 335 data with the composition of 85 entrepreneurs and 250 non-entrepreneurs (a ratio of approximately 1:3). Since the data condition had not yet reached the desired balance target, the under-sampling technique was applied using ten different methods, as shown in Table 2.

Table 2. Data comparison after under sampling

Under sampling methods	Entrepreneur	Non-entrepreneur	Ratio
Cluster Centroids (CC)	85	170	1 : 2.00
Condensed Nearest Neighbour (CNN)	85	75	1 : 0.88
Edited Nearest Neighbours (ENN)	85	171	1 : 2.01
Repeated Edited Nearest Neighbours (RENN)	85	151	1 : 1.78
All k-Nearest Neighbour (AllKNN)	85	158	1 : 1.86
Instance Hardness Threshold (IHT)	85	176	1 : 2.07
Near Miss (NM)	85	170	1 : 2.00
One Sided Selection (OSS)	85	202	1 : 2.38
Random Under Sampling (RUS)	85	170	1 : 2.00
Tomek Links (TL)	85	230	1 : 2.71

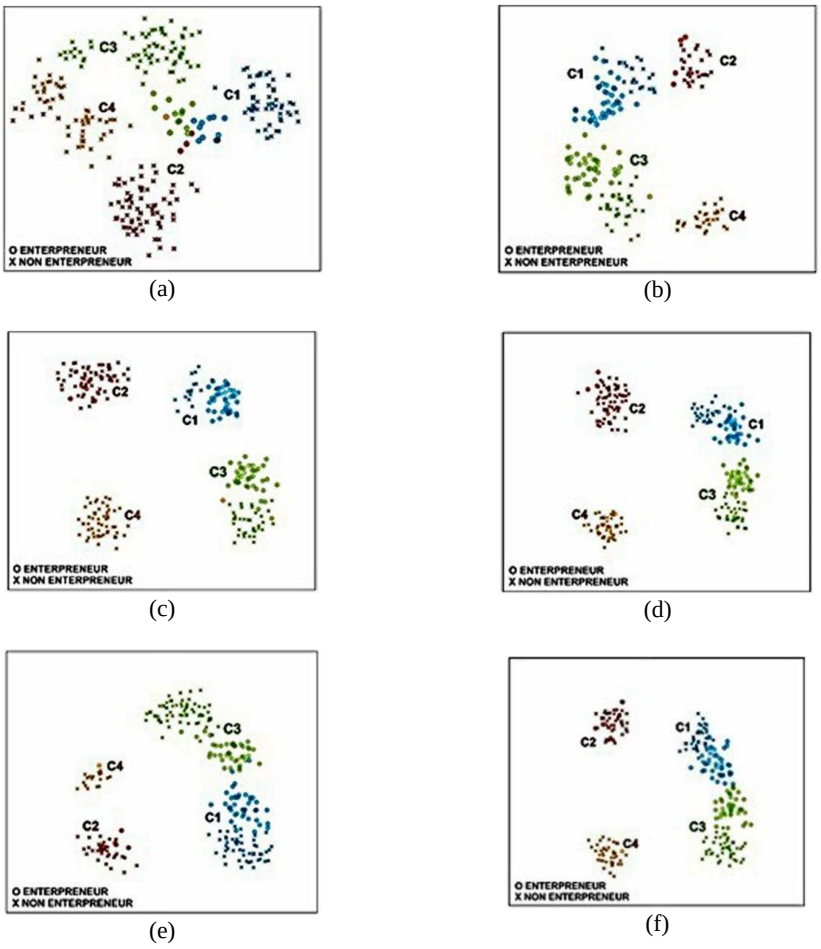

Fig. 2. *t*-SNE visualizations of four clusters: without sampling (a); after CC sampling (b); after ENN sampling (c); after IHT sampling (d); after NM sampling (e); after RUS sampling (f)

Some under-sampling techniques, CC, ENN, IHT, NM, and RUS, were successful in balancing non-entrepreneur data and reached a ratio of 1:2 between entrepreneurs and non-entrepreneur. Fig. 4 visualizes the data clusters using the k-Means method on the balanced dataset. Some data appear to be utterly separated according to their respective clusters, although some data are mixed with other

clusters (Fig. 4a, 4e, 4f). The sampling results of ENN (Fig. 4c) and IHT (Fig. 4d) produced four entirely separated clusters. The better quality of over-sampling SMOTE-NC and under-sampling ENN and IHT in balancing small datasets is noticeable in the following experiment for classifying the statuses of entrepreneurs and non-entrepreneur with different datasets.

4.4. Attribute selection and verification of sampling techniques

Although the dataset has been balanced after the sampling process, a step of the attribute selection process is carried out to determine the most important attributes among TPB variables that affected the entrepreneurial status of students (entrepreneur/non-entrepreneur). The attribute selection process was performed using Information Gain (IG), and the five attributes with the highest IG values are Att (A3), SN (A6, A10, A11), and PBC (A13). Then, we have five datasets of Data01 (A1-A14), Data02 (A3, A6, A10, A11, A13), Data03 (A3, A6, A11, A13), Data04 (A3, A6, A11, A13, A14), and Data05 (A2, A3, A6, A11, A13). Attribute A10 had a higher rank value than the other four A3, A6, A11, and A13. Thus, Data03 was formed by eliminating the dominant attribute. Data04 and Data05 were designed by replacing them with other TPB variables like PBC of A4 and Att of A2. We designed different datasets to observe the TPB variables' combination to describe the student entrepreneurial potential better.

We experimented with different datasets with REPTree, J48, OneR, PART, RIPPER classifiers. Then made some randomization testing using the Stratified Cross-Validation technique with ten folds. The average accuracy performances (Area Under the Curve (AUC)) from those five classifiers using three datasets of Data01, Data02, and Data03 can be seen in Fig. 3.

Higher AUC values, 0.96, 0.86, and 0.86 respectively, for those three datasets, were obtained from two sampling methods of SMOTE-NC+IHT and SMOTE-NC+ENN, which are consistent with previous sampling studies on small datasets [22]. The results showed that the two sampling techniques overcame the imbalanced data and substantially impacted the classification performance. Although AUC is analogous to an accuracy indicator in gauging a model, AUC has been stated to represent the unbalanced problem better.

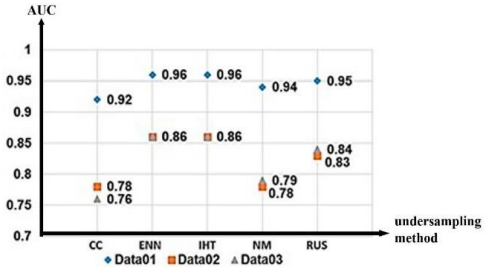

Fig. 3. AUC average values with various under-sampling scenarios

4.5. Generating rules for detecting student entrepreneurial potential

This stage aims to obtain rules on five datasets (Data01-Data05) to detect the entrepreneurial potential of students. This is done using rule-based algorithms,

namely PART, OneR, and RIPPER, and decision tree techniques such as J48 and REPTree. Data01 is the original data with all TPB variables. Data02 only utilizes the selected variables, while Data03-Data05 are the variation of Data02 as mentioned in the preceding section. Although these five datasets have different attributes, they still contain at least one representation of TPB variables for Att, SN, and PBC. The classification experiment recorded the average performance of those five datasets to show the effect of composition on the stability of TPB variables with different sampling scenarios. The experimental results in Fig. 4 illustrate the average classification performance of five datasets with REPTree, J48, OneR, PART, and RIPPER, which are presented in the spider chart for three sampling scenarios – without sampling (a), as well as sampled by SMOTE-NC+IHT (b) and SMOTE-NC+ENN (c) methods.

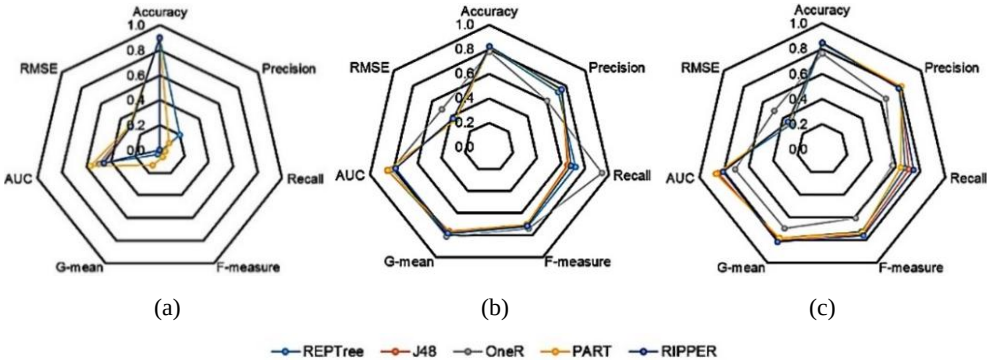

Fig. 4. The average performance of classification algorithm on: original dataset (a); with Smote-NC+ENN dataset (b); and with Smote-NC+IHT dataset (c)

Fig. 4a shows relatively higher accuracy and AUC values in addition to a lower error value of RMSE. However, even the models detect some students with entrepreneurial potential. They might fail to retrieve other students with nearly similar characteristics, signified by lower precision, recall, and F-measure values. Because the first scenario used un-sampled data, the Geometric Mean (G-Mean) that measures the balance between classification performances of majority and minority classes also has a lower value. Thus, the models of five datasets with an un-sampled approach could not be considered a solution for detecting student entrepreneurial status.

Fig. 4b and 4c demonstrate better classification models with the sampling scenario. The evidence of their performances is relatively high average values on those various indicators of accuracy and AUC and the goodness indicators in detecting more students with entrepreneurial interest like precision, recall, and their balancing ratio in F-measure and G-Mean. The models also have lower error values of RMSEs. These findings suggest that the algorithms could classify the majority and minority classes and obtain genuine positive data by labelling them positive. They can do the same for data with negative values. However, since there was a model of OneR on Smote-NC+ENN sampling scenario that had relatively higher values of recall and RMSE, it could be stated that the third sampling scenario on Smote-NC+IHT is better than the second one. A better classification model should have

higher precision and lower recall. This finding shows that the dataset sampled with SMOTE-NC+ENN could produce instability in the classification algorithm's performance.

Table 3. Results for experiments using various rule-based algorithms

Algo-rithm	Dataset	Performance						
		Accuracy	Precision	Recall	F-measure	G-mean	AUC	RMSE
REPTree	Data01	0.893	0.843	0.824	0.833	0.873	0.906	0.298
	Data02	0.885	0.831	0.812	0.821	0.864	0.883	0.310
	Data03	0.793	0.754	0.541	0.630	0.704	0.833	0.266
	Data04	0.824	0.800	0.612	0.693	0,753	0.859	0.358
	Data05	0.785	0.704	0,588	0.641	0.720	0.830	0.375
	Average	*0.836*	*0.786*	*0.675*	*0.724*	*0.783*	*0.862*	*0.321*
J48	**Data01**	0.900	0.864	0.824	0.843	0.879	0.833	0.308
	Data02	0.896	0.837	0.847	0.842	0.883	0.883	0.301
	Data03	0.804	0.736	0.624	0.675	0.746	0.844	0.361
	Data04	0.824	0.800	0.612	0.693	0.753	0.839	0.364
	Data05	0.801	0.726	0.624	0.671	0.743	0.842	0.364
	Average	**0.845**	**0.793**	**0.706**	**0.745**	**0.801**	**0.848**	**0.340**
OneR	Data01	0.736	0.600	0.565	0.582	0.680	0.691	0.514
	Data02	0.736	0.600	0.565	0.582	0.680	0.691	0.514
	Data03	0.774	0.681	0.576	0.624	0.708	0.723	0.476
	Data04	0.774	0.681	0.576	0.624	0.708	0.723	0.475
	Data05	0.774	0.681	0.576	0.624	0.708	0.723	0.476
	Average	*0.759*	*0.649*	*0.572*	*0.607*	*0.697*	*0.710*	*0.491*
PART	Data01	0.885	0.831	0.812	0.821	0.864	0.871	0.317
	Data02	0.896	0.854	0.824	0.838	0.876	0.898	0.304
	Data03	0.793	0.782	0.506	0.614	0.686	0.849	0.367
	Data04	0.854	0.862	0.659	0.747	0.791	0.879	0.341
	Data05	0.774	0.686	0.565	0.619	0.703	0.841	0.384
	Average	*0.840*	*0.803*	*0.639*	*0.728*	*0.784*	*0.868*	*0.343*
RIPPER	Data01	0.881	0.793	0.859	0.825	0.875	0.884	0.316
	Data02	0.896	0.837	0.847	0.842	0.883	0.848	0.305
	Data03	0.812	0.737	0.659	0.696	0.764	0.754	0.385
	Data04	0.839	0.795	0.682	0.734	0.790	0.786	0.370
	Data05	0.812	0.737	0.659	0.696	0.764	0.778	0.382
	Average	*0.848*	*0.780*	*0.741*	*0.759*	*0.815*	*0.810*	*0.352*

Thus, the scheme sampled by SMOTE-NC + IHT was deemed worth considering as a solution to detect student entrepreneurial status. The performances of the five classification algorithms are listed in Table 3. The J48 algorithm's performance appears to be the most stable with relatively better average values in all indicators of accuracy, precision, recall, F-measure, G-mean, AUC, and RMSE. Its results are more consistent across all datasets than other algorithms, with the best performance in Data01 (all attributes) and Data02 (the top five selected attributes). Based on the AUC values, the J48 classification results of those two datasets are defined in the Good Classification category. It produces almost the same performance in classifying 2 data classes of 85 entrepreneurs and 176 non-entrepreneurs, as represented by the confusion matrix in Table 4.

The J48 algorithm of Data01 with an overall accuracy average of 90% (Table 3) was unsuccessful in classifying 15 students for having entrepreneur potential since they were declared as non-entrepreneurs (with recall 82.4% in Table 4). Meanwhile, in Data02, the algorithm results have a lower average classification accuracy of 89.7% (Table 3) but a higher recall value (84.7% in Table 4). Thus, the decision tree

formed by J48 is extracted into IF-Then rules for detecting student entrepreneur status as listed in Table 5.

Table 4. Confusion matrix for J48 classification results using SMOTE-NC+IHT sampling method on certain datasets

Entrepreneur potential	Data01			Data02		
	Predicted entrepreneur	Predicted not entrepreneur	Recall	Predicted entrepreneur	Predicted not entrepreneur	Recall
Actual entrepreneur	70	15	82.4%	72	13	84.7%
Actual not entrepreneur	11	165	93.8%	14	162	92.0%
Precision	86.4%	91.7%		83.7%	92.6%	

Table 5. Rules for detecting student entrepreneurial status for Data01

Rule R1	**IF** (A13=nonmember) **THEN** *non-entrepreneur*
Rule R2	**IF** (A13=member) AND (A8=(C or D)) **THEN** *entrepreneur*
Rule R3	**IF** (A13=member) and (A8=B) and (A14= (P or F)) and (A11= (B or C)) and (A1=NonSch) **THEN** *entrepreneur*
Rule R4	**IF** (A13=member) and (A8=B) and (A14= (P or F)) and (A11= (B or C)) and (A1=Sch) and (A7= (B or C or E)) **THEN** *entrepreneur*
Rule R5	**IF** (A13=member) and (A8=B) and (A14= (P or F)) and (A11= (B or C)) and (A1=Sch) and (A7= A) **THEN** *non-entrepreneur*
Rule R6	**IF** (A13=member) and (A8=B) and (A14= (P or F)) and (A11=A) and (A10=C) **THEN** *non-entrepreneur*
Rule R7	**IF** (A13=member) and (A8=B) and (A14= (P or F)) and (A11=A) and (A10= (A or B)) and (A5=P5) **THEN** *non-entrepreneur*
Rule R8	**IF** (A13=member) and (A8=B) and (A14= (P or F)) and (A11=A) and (A10= (A or B)) and (A5= (P2 or P3 or P4)) **THEN** *entrepreneur*
Rule R9	**IF** (A13=member) and (A8=A) and (A14= (P or F)) and (A10= (B or C)) **THEN** *entrepreneur*
Rule R10	**IF** (A13=member) and (A8=A) and (A14= (P or F)) and (A10=A) and (A5= (P4 or P5)) and (A2=NonAct) **THEN** *non-entrepreneur*
Rule R11	**IF** (A13=member) and (A8=A) and (A14= (P or F)) and (A10=A) and (A5= (P5 or P4)) and (A2=Act) **THEN** *entrepreneur*
Rule R12	**IF** (A13=member) and (A8=A) and (A14= (P or F)) and (A10=A) and (A5= (P3 or P2)) **THEN** *entrepreneur*
Rule R13	**IF** (A13=member) and (A8= (A or B)) and (A14=E), **THEN** *entrepreneur*

13 rules to detect student entrepreneurial status have been generated from the J48+(SMOTE-NC+IHT) scheme applied to the Data01 dataset (Table 5). TPB variable composition of Att, SN, and PBC in Table 6, i.e., eight rules are for entrepreneur status, and five rules are for non-entrepreneurs status. Attribute A13 forms the root of the student entrepreneurial status decision tree in all rules. For the shortest path (R1), the decision tree only involves the PBC of A13, which indicates that the PBC variable could directly influence behavior [15] and is the essential variable in the rule base for entrepreneurial status. As the path lengthens, the SN variable position closer to the root provides opportunities for universities to play a more significant role through effective curriculum design as course activities for strengthening individual perceptions via vertical relationships between lecturer-student. Meanwhile, the Att variable, which is relatively far from the root, shows that students' individual beliefs could still be influenced by the SN factor closer to the root.

Table 6. TPB variables on the detecting rules of student entrepreneur potential status in Data01 and their performances

Status	Rule	Att	SN	PBC	Coverage	Accuracy
Entrepreneur	R2	-	A8	A13	8.43%	100.00%
	R3	A1	A8, A11	A13, A14	4.21%	100.00%
	R4	A1	A8, A11, A7	A13, A14	1.53%	75.00%
	R8	-	A8, A11, A10, A5	A13, A14	3.45%	88.89%
	R9	-	A8, A10	A13, A14	5.36%	100.00%
	R11	A2	A8, A10, A5	A13, A14	0.77%	100.00%
	R12	-	A8, A10, A5	A13, A14	1.53%	100.00%
	R13	-	A8	A13, A14	4.98%	100.00%
Result					*Total: 30.26%*	*Average: 95.49%*
Non-entrepreneur	R1	-	-	A13	46.36%	95.87%
	R5	A1	A8, A11, A7	A13, A14	1.53%	100.00%
	R6	-	A8, A11, A10	A13, A14	18.77%	93.88%
	R7	-	A8, A11, A10, A5	A13, A14	1.15%	100.00%
	R10	A2	A8, A10, A5	A13, A14	1.93%	100.00%
Result					*Total: 69.74%*	*Average: 97.95%*

Some of the important rule patterns that formed the decision tree are:

a. If the student was a non-participant of the Business Incubator, they were automatically classified as a non-entrepreneur (R1).

b. Students who were participants of the Business Incubator had an Excellent status CSEP and passed A8 courses were automatically classified as entrepreneurs (R2 and R13).

c. A student who was a Business Incubator had a Proposing or Funded CSEP status either as a scholarship recipient or not and passed the A7, and A11 courses was automatically an entrepreneur (R3, R4). However, if they were a scholarship recipient whose course score was A7 = A, the recipient would likely become a non-entrepreneur (R5).

d. Students who were participants of the Business Incubator had Proposing or Funded CSEP status, passed A10 courses with their parents' income <5 million, and registered as activists would likely become entrepreneurs (R11). However, if they were not activists, they were classified as non-entrepreneurs (R10).

e. Students who become the participants of the Business Incubator, had Proposing or Funded CSEP status, passed A10, and A11 courses with their parents' income reaching more than 10 million were classified as non-entrepreneurs (R7). However, if their parents' income was less than 10 million, the rules detect them as entrepreneurs (R8 and R12).

The points above show that the SN variable's relationship between attribute values and students' entrepreneurial statuses is not linear, i.e., the A7 usage in R3, R4, and R5 could have both labels as entrepreneurs and non-entrepreneurs. This condition indicates the need to review the content and outcomes of subjects related to entrepreneurship to reflect a linear relationship with students' entrepreneurial behavior.

Table 7. Rules for detecting student entrepreneurial status for Data02

Rule R1	**IF** (A13=non-member) **THEN** *Non-entrepreneur*
Rule R2	**IF** (A13=member) and (A10=C) and (A11= (B or C)) **THEN** *Entrepreneur*
Rule R3	**IF** (A13=member) and (A10=C) and (A11=A) **THEN** *Non-entrepreneur*
Rule R4	**IF** (A13=member) and (A10= (A or B)) **THEN** *Entrepreneur*

Table 8. TPB variables on the detecting rules of student entrepreneur potential status in Data02 and their performances

Status	Rule	Att	SN	PBC	Coverage	Accuracy
Entrepreneur	R2	-	A10, A11	A13	7.66%	95.00%
	R4	-	A10	A13	25.29%	80.30%
Result					*Total: 32.95%*	*Average: 87.65%*
Non-entrepreneur	R1	-	-	A13	46.36%	95.87%
	R3	-	A10, A11	A13	20.69%	85.19%
Result					*Total: 67.05%*	*Average: 90.53%*

The J48 + (SMOTE-NC+IHT) scheme applied to the Data02 dataset produced four rules to identify the student entrepreneur status (Table 7) with the composition of TPB variables in Table 8. Although the generated rules only involved two TPB variables of SN and PBC, this ruleset is still worth considering based on the TPB concept in Fig. 1 [15]. In this ruleset, A13 representing the PBC variable is an important attribute and functioned as the root for forming rules for detecting student entrepreneurial status. Some of those rules are translated as follows.

a. Rule R1 indicates that if a student were not in a Business Incubator, they could automatically believe that they were not entrepreneurs. However, if a student was a Business Incubator member, he/she could become an entrepreneur, except when the attribute values were A10=C and A11=A.

b. This finding indicates that university business incubator activities can positively influence students on performing a specific entrepreneurship behavior. Business Incubators could grow self-efficacy, namely the belief in an individual student that he/she could acquire the skills needed as an entrepreneur.

c. In rule R3, the attribute values of A10=C and A11=A, in conjunction with the value A13=member, would classify student entrepreneurial status as a non-entrepreneur.

d. Meanwhile, in rules R2 and R4, the value variants of A10 and A11 in conjunction with value A13=member classifies student entrepreneurial status as an entrepreneur. The A13 value also influences the role SN variables of A10 and A11. The finding strengthens the indication that the PBC variable is the essential factor in the rule base for entrepreneur status.

Thus, the experiments have successfully initiated the algorithms to generate detection rules that are adaptive to the dataset conditions.

4.6. Performance Evaluation on Generated Ruleset

The performance of the entrepreneur's status detection ruleset was evaluated based on its coverage and accuracy. The J48 + (SMOTE-NC+IHT) scheme on Data01, which has 261 data (see Table 2), successfully classified 30.26% of data members as having the status to become entrepreneurs with the coverage value of each rule listed in Table 6. The results display the lowest coverage value of 0.77% in rule R11, which

means two data are included as entrepreneurs. A rule is defined as exhaustive coverage to detect every possible combination of attribute values. From Table 2, the data ratio after sampling methods for the entrepreneur: non-entrepreneurs is 1:2. Thus, the coverage values of rules to detect the entrepreneur status as 30.26% of data members indicated a satisfactory ruleset.

Additionally, the ruleset to classify non-entrepreneur status were also successful in detecting 69.74% of data members, with the highest coverage value being 46.36%. From 261 data members in Data01, 121 data (46.36%) were detected through Rule R1. Aside from coverage values, Table 6 shows the 100% accuracy for six rules of detecting the entrepreneur status with an average of 95.49%. In comparison, there were only three rules of non-entrepreneurs with 100% accuracy and the average accuracy of all rules being 97.95%. Rule R1 had an accuracy of 95.87% which means from 121 non-entrepreneur data members, 116 students were correctly classified. Alternatively, R2 became the preferable rule for detecting the entrepreneur status.

Table 8 shows the Data02 results with the Smote-NC+IHT sampling technique related to the ruleset generated from J48 and its TPB variables and the values of coverage and accuracy for each rule. The average accuracy of all rules detecting the status as an entrepreneur was 87.65% and 90.53% for the non-entrepreneur. Data01 and Data02 have 261 data members since both datasets have been balanced with the similar sampling method of Smote-NC+IHT. However, they have different attribute values. Because of those differences in which Data02 only considers the top five selected attributes, the coverage values of Data02 to detect entrepreneurial status as minority class are higher with 32.95% in total.

Based on the coverage value and accuracy of the rules generated in Data01 and Data02, each ruleset has advantages and disadvantages. The ruleset of non-entrepreneurs from Data01 produces higher accuracy than from Data02. The accuracy of 97.95% compared to 90.53% is, with a difference of 7.42%. However, the composition of the coverage ruleset of both datasets is more similar for detecting a student's status as an entrepreneur or non-entrepreneurs, respectively, 30.26%: 69.74% for Data01 and 32.95%: 67.05% for Data02. Therefore, there is an error possibility of ± 2% for detection rules of entrepreneur status. Although the ruleset of Data02 with five attributes has lower accuracy, they have higher coverage values that makes the ruleset have a lower potential for data error in each detection rule for entrepreneurial status. The facts generated in this experiment pose an option for applying rulesets to machine learning systems by considering the conditions of a dataset.

5. Discussions

The experimental results indicate that university databases could be used to explore the individual behavioral patterns of students using the TPB approach. However, the student records require preprocessing before becoming viable research datasets. An imbalanced class handling process and standard data cleaning and integration preprocessing were required because the data availability related to student entrepreneurial status was minimal. Experimental sampling techniques were carried

out in three different dataset scenarios by adding the NC feature to the SMOTE [20] and combining it with IHT or ENN under-sampling technique to balance the dataset. Datasets balanced by SMOTE-NC+IHT and SMOTE-NC+ENN and clustered with k-Means have produced better representation. These two sampling schemes also consistently produced higher average AUC values for various classifiers of REPTree, J48, OneR, PART, and RIPPER. In the context of behavior, it could be interpreted that the dataset balanced with SMOTE-NC+IHT [22] or SMOTE-NC+ENN techniques consisted of groups of individuals who had similar characteristics in their group members, as well as differences in characteristics with other group members. Thus, the dataset was considered feasible to train machine learning systems to form more precise data classes.

After forming datasets from sampling methods, our experiments гьэе observed some variants of datasets with classification methods. The results showed that the combination of Smote-NC+IHT and the rule generator algorithm with the J48 decision tree on the 14-attribute and the 5-attribute datasets had produced rulesets with a lower potential for data error. Rulesets using 14 attributes based on the root of the PBC variable lead to an average accuracy rate of 97.95% and an average potential data error of ±2% in detecting entrepreneurial status. The ruleset showed a non-linear relationship between the SN variable's attribute values and student entrepreneurial status. This condition indicates the ineffective impact of the courses generated on the dataset in shaping student's entrepreneurial behavior. Similar findings for the rulesets from Data01 were found in the rulesets from Data02, but with a lower potential for data error since they have larger coverage value in the minority class of entrepreneur status.

In general, the two rulesets use the same reasoning to detect student entrepreneurial status. These two rulesets are worthy of consideration, as they both place the PBC variable (A13) as the root that forms detection rules by the TPB concept. They also confirm that PBC variables could directly determine behavior. The first ruleset of three TPB variables from Data01 have more significant potential for data error than the second ruleset of two TPB variables from Data02. The performance difference of the two rulesets trained with SMOTE-NC+ IHT and J48 is insignificant. The second ruleset proves that the selected attribute dataset could generate detection rules with a performance different from the ones generated from the dataset for which attribute selection has not been performed [11]. This indicates that the proposed model is quite adaptive, even when applied to two types of datasets that significantly differed in their number of attributes. This proposed model supports long-term comprehension and planning based on academic information by focusing on behavioral variables in TPB. Therefore, by considering dataset conditions, these two alternative rulesets are options for ruleset implementation in machine learning.

6. Conclusions and future work

This study produces a model for early detection of student entrepreneurial status in two sets of rules generated from two different dataset variants. The first model uses a dataset of fourteen attributes and produces thirteen rules with an accuracy rate of

97.95% and a potential error of 2.31%. The second model uses a dataset of five attributes and produces four rules with an average accuracy rate of 90.53%, but the potential error is 0.38%. The two models use identical reasoning in detecting the entrepreneurial status of students, namely using the PBC variable as the root of the decision tree structure. Two sets of detection rules reveal two SN variable representation attributes to support the attributes of the business incubator status as PBC variable representations. Professional Ethics and Career Guidance courses have a powerful influence in shaping student entrepreneurial behavior. In a more complex ruleset (13 rules), the significance of the influence of these two variables is increasingly visible with the growth of branches on the decision tree based on the leaf node attributes of PKM Status (PBC variable) and English (SN variable).

Further development of the detection model can be done by exploring preprocessing techniques using high-dimensional datasets with more instances with a more complex composition of TPB attributes. Experiments were also conducted to measure the effect of TPB attributes more specifically on various variants of more complex datasets to reveal student entrepreneurial behavior more comprehensively.

References

1. Ajzen, I. The Theory of Planned Behavior. – Organ. Behav. Hum. Decis. Process., Vol. **50**, 1991, No 2, pp. 179-211. DOI: https://doi.org/10.1016/0749-5978(91)90020-T.
2. Sánchez, V. B., C. A. Sahuquillo. Entrepreneurial Intention Among Engineering Students: The Role of Entrepreneurship Education. – Eur. Res. Manag. Bus. Econ., Vol. **24**, 2018, No 1, pp. 53-61. DOI: https://doi.org/10.1016/j.iedeen.2017.04.001.
3. Botsaris, C., V. Vamvaka. Attitude Toward Entrepreneurship: Structure, Prediction from Behavioral Beliefs, and Relation to Entrepreneurial Intention. – J. Knowl. Econ., Vol. **7**, 2016, No 2, pp. 433-460. DOI: 10.1007/s13132-014-0227-2.
4. Tsordia, C., D. Papadimitriou. The Role of Theory of Planned Behavior on Entrepreneurial Intention of Greek Business Students. – Int. J. Synerg. Res., Vol. **4**, 2015.
5. Utami, C. W. Attitude, Subjective Norms, Perceived Behavior, Entrepreneurship Education and Self-Efficacy toward Entrepreneurial Intention University Student in Indonesia. – Eur. Res. Stud. J., Vol. **20**, 2017, pp. 475-495.
6. Karimi, S., H. Biemans, K. Mahdei, T. Lans, M. Chizari, M. Mulder. Testing the Relationship between Personality Characteristics, Contextual Factors and Entrepreneurial Intentions in a Developing Country. – Int. J. Psychol., Vol. **52**, 2015. DOI: 10.1002/ijop.12209.
7. Maresch, D., R. Harms, N. Kailer, B. Wimmer-Wurm. The Impact of Entrepreneurship Education on the Entrepreneurial Intention of Students in Science and Engineering Versus Business Studies University Programs. – Technol. Forecast. Soc. Change, Vol. **104**, 2016, pp. 172-179. DOI: https://doi.org/10.1016/j.techfore.2015.11.006.
8. Hongyi, S., L. C. Tung, L. Bo, W. Y. L. Belle. The Impact of Entrepreneurial Education on Entrepreneurial Intention of Engineering Students in Hong Kong. – Manag. Decis., Vol. **55**, January 2017, No 7, pp. 1371-1393. DOI: 10.1108/MD-06-2016-0392.
9. Rijati, N., D. Purwitasari, S. Sumpeno, M. H. Purnomo. A Decision Making and Clustering Method Integration Based on the Theory of Planned Behavior for Student Entrepreneurial Potential Mapping in Indonesia. – Int. J. Intell. Eng. Syst., Vol. **13**, 2020, No 4. DOI: 10.22266/ijies2020.0831.12.
10. Thammasiri, D., D. Delen, P. Meesad, N. Kasap. A Critical Assessment of Imbalanced Class Distribution Problem: The Case of Predicting Freshmen Student Attrition. – Expert Syst. Appl., Vol. **41**, 2014, No 2, pp. 321-330. DOI: 10.1016/j.eswa.2013.07.046.

11. M á r q u e z-V e r a, C., A. C a n o, C. R o m e r o, S. V e n t u r a. Predicting Student Failure at School Using Genetic Programming and Different Data Mining Approaches with High Dimensional and Imbalanced Data. – Appl. Intell., Vol. **38**, 2013, No 3, pp. 315-330. DOI: 10.1007/s10489-012-0374-8.
12. M a t h e w s, L. M., H. S e e t h a. On Improving the Classification of Imbalanced Data. – Cybernetics and Information Technologies, Vol. **17**, 2017, No 1, pp. 45-62.
13. X i a o, J., L. X i e, C. H e, X. J i a n g. Dynamic Classifier Ensemble Model for Customer Classification with Imbalanced Class Distribution. – Expert Syst. Appl., Vol. **39**, 2012, No 3, pp. 3668-3675. DOI: https://doi.org/10.1016/j.eswa.2011.09.059.
14. E l r e e d y, D., A. F. A t i y a. A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for Handling Class Imbalance. – Inf. Sci., Vol. **505**, 2019, pp. 32-64. DOI: https://doi.org/10.1016/j.ins.2019.07.070.
15. A j z e n, I., M. F i s h b e i n. The Influence of Attitudes on Behavior. – In: The Handbook of Attitudes. Vol. **173**. 2005, pp. 173-221.
16. K h e f a c h a, I., L. B e l k a c e m. Modeling Entrepreneurial Decision-Making Process Using Concepts from Fuzzy Set Theory. – J. Glob. Entrep. Res., Vol. **5**, 2015, No 1, p. 13. DOI: 10.1186/s40497-015-0031-x.
17. S h o v o n, M. H. I., M. H a q u e. An Approach of Improving Student's Academic Performance by Using k-Means Clustering Algorithm and Decision Tree. – Int. J. Adv. Comput. Sci. Appl., Vol. **3**, 2012. DOI: 10.14569/IJACSA.2012.030824.
18. de M o r a i s, A. M., J. M. F. R. A r a ú j o, E. B. C o s t a. Monitoring Student Performance Using Data Clustering and Predictive Modelling. – In: Proc. of IEEE Frontiers in Education Conference (FIE) Proceedings, Oct. 2014, pp. 1-8, DOI: 10.1109/FIE.2014.7044401.
19. R i j a t i, N., S. S u m p e n o, M. H. P u r n o m o. Multi-Attribute Clustering of Student's Entrepreneurial Potential Mapping Based on Its Characteristics and the Affecting Factors: Preliminary Study on Indonesian Higher Education Database. – In: Proc. of 10th International Conference on Computer and Automation Engineering, 2018, pp. 11-16. DOI: 10.1145/3192975.3193014.
20. C h a w l a, N. V., K. W. B o w y e r, L. O. H a l l, W. P. K e g e l m e y e r. SMOTE: Synthetic Minority Over-Sampling Technique. – J. Artif. Intell. Res., Vol. **16**, 2002, No February 2017, pp. 321-357. DOI: 10.1613/jair.953.
21. L e m a i t r e, G., F. N o g u e i r a, C. K. A r i d a s. Imbalanced-Learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. – J. Mach. Learn. Res., Vol. **18**, January 2017, No 1, pp. 559-563.
22. B a t i s t a, G. E. A. P. A., R. C. P r a t i, M. C. M o n a r d. A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data. – SIGKDD Explor. Newsl., Vol. **6**, 2004, No 1, pp. 20-29. DOI: 10.1145/1007730.1007735.
23. W i l s o n, D. L. Asymptotic Properties of Nearest Neighbor Rules Using Edited Data. – IEEE Trans. Syst. Man Cybern., Vol. **2**, 1972, No 3, pp. 408-421. DOI: 10.1109/TSMC.1972.4309137.
24. T o m e k, I. An Experiment with the Edited Nearest-Neighbor Rule. – IEEE Trans. Syst. Man. Cybern., Vol. **SMC-6**, 1976, No 6, pp. 448-452. DOI: 10.1109/TSMC.1976.4309523.
25. G o w d a, K., G. K r i s h n a. The Condensed Nearest Neighbor Rule Using the Concept of Mutual Nearest Neighborhood (Corresp.). – IEEE Trans. Inf. Theory, Vol. **25**, Jul. 1979, No 4, pp. 488-490. DOI: 10.1109/TIT.1979.1056066.
26. K u b a t, M. Addressing the Curse of Imbalanced Training Sets: One-Sided Selection. – In: Proc. of 14th Int. Conf. Mach. Learn., 2000.
27. S m i t h, M. R., T. M a r t i n e z, C. G i r a u d-C a r r i e r. An Instance Level Analysis of Data Complexity. – Mach. Learn., Vol. **95**, 2014, No 2, pp. 225-256. DOI: 10.1007/s10994-013-5422-z.
28. R i j a t i, N., S. S u m p e n o, M. H. P u r n o m o. Attribute Selection Techniques to Clustering the Entrepreneurial Potential of Student Based on Academic Behavior. – In: Proc. of IEEE Int. Conf. Comput. Intell. Virtual Environ. Meas. Syst. Appl. CIVEMSA 2019 – Proc., 2019. DOI: 10.1109/CIVEMSA45640.2019.9071597.
29. Y u h a n a, U., et al. A Rule-Based Expert System for Automatic Question Classification in Mathematics Adaptive Assessment on Indonesian Elementary School Environment. – Int. J. Innov. Comput. Inf. Control, Vol. **15**, 2019, pp. 143-161. DOI: 10.24507/ijicic.15.01.143.

30. P r a c h u a b s u p a k i j, W., P. D o u n g p a i s a n. Matching Preprocessing Methods for Improving the Prediction of Student's Graduation. – In: Proc. of 2nd IEEE International Conference on Computer and Communications (ICCC), Oct. 2016, pp. 33-37. DOI: 10.1109/CompComm.2016.7924659.

Received: 19.10.2021; Second Version: 15.02.2022; Accepted: 12.03.2022