Turkish Journal of Forecasting vol. 06 no. 2 (2022) pp. 53-60

Content list available at JournalPark

Turkish Journal of Forecasting

Journal Homepage: tjforecasting.com

The Effect of Handling Imbalanced Datasets Methods on Prediction of Entrepreneurial Competency in University Students

M. Simsek[1,*], A.S. Das[2]

[1]*Ostim Technical University, Faculty of Engineering, Department of Artificial Intelligence Engineering, Ankara, Turkey*
[2]*Ozyegin University, Faculty of Engineering, Department of Industrial Engineering, Istanbul, Turkey*

A R T I C L E I N F O

Article history:

Received	07	October	2022
Revision	31	October	2022
Accepted	01	November	2022
Available online	31	December	2022

Keywords:

Machine learning
Imbalance data
Handling imbalanced dataset methods
Entrepreneurs
Classification

R E S E A R C H A R T I C L E

A B S T R A C T

As of today, entrepreneurs and entrepreneurship are considered to be the integral parts of the economic and technological advancements. Entrepreneurs are promoted in many countries because of their high return on investment opportunities both in terms of income and new inventions. Numerous studies prove that entrepreneurs have many traits in common and these common traits can correlate with each other. Based on these common traits, potential entrepreneurs can be predicted, current entrepreneurs can be improved by realising their weak sides and the ones who wish to be entrepreneurs can be provided with insights. A machine learning approach can light the way for a better rewarding future for entrepreneurship, helping these goals significantly. There exist several studies for the prediction of entrepreneurial competency with the use of machine learning algorithms. Most machine learning methods perform better accuracy and F1-score imbalanced data instead in imbalanced data. This study focuses on utilizing imbalanced class handling methods to increase prediction performance. Random Oversampling, Random Undersampling, SMOTE, and NearMiss methods are used to handling imbalanced data for this purpose in this study. The performance of the machine learning algorithms with Imbalanced Data Handling methods is compared with the machine learning algorithms without these methods. The comparison shows that with the handling imbalanced data methods machine learning algorithms perform better.

1. Introduction

Entrepreneurship is strongly linked to the levels of economic development [1]. Entrepreneurs are supported and rewarded by governments around the world. Furthermore, the individual desire to be an entrepreneur is also noticeably high worldwide. In many countries, more than half of all individuals reported a desire for self-employment [2].

There are various studies [3, 4] about the common traits of entrepreneurs. Today, it is supported by more than enough empirical evidence that entrepreneurs statistically have multiple traits in common [5]. Many studies demonstrate that the distribution of common traits may differ for a vast number of conditions

*Corresponding author.
E-mail addresses: murat.simsek@ostimteknik.edu.tr (Murat Şimşek)

https://doi.org/10.34110/forecasting.1185545

including different sectors [6], cultures [7], economic environments [8], and gender [9,10]. Different conditions cause different key traits for entrepreneurial success, causing countless combinations. For this reason, the study of entrepreneurial competency and common traits of entrepreneurs is suggested to be handled by machine learning methods.

There have been several pieces of research for predicting entrepreneurial competency using machine learning algorithms [10-12]. This study aims to achieve higher performance on base machine learning algorithms by applying imbalanced data handling techniques. In order to observe the effect of imbalanced data handling techniques, base machine learning methods and machine learning methods used together with imbalance data handling techniques are compared with comparison parameters.

2. Materials and Methods

2.1. The dataset

The dataset used for this analysis is about entrepreneurial competency in university students [12-13]. The data provided a set of questions to 219 Indian university students to determine their entrepreneurial competency [12]. Students were surveyed to assess their traits and state their backgrounds. Our study uses all but one of the questions from the data. The question "Reason for lack" was left out intentionally because every student did not answer the question and the answers were commentary; therefore, it could not be converted into a numeric attribute for the machine learning algorithm.
Original dataset contains 16 columns, it is worth noting that the question labelled as “mental disorder” is the original question to students about their mental state in terms of stress.

The correlation between features is given in a heatmap of Pearson’s correlation matrix. It is shown in Figure 1 that the correlations between features are not high enough to make eliminations.

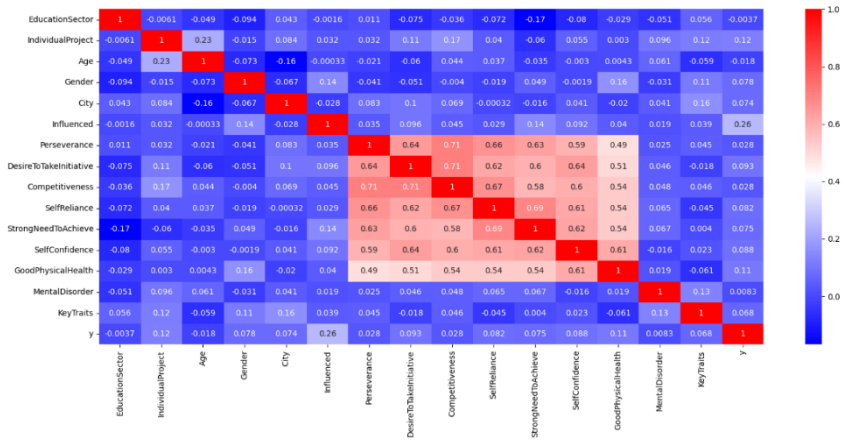

Figure 1. Correlation between features

2.2. Classification algorithms

The prediction phase of this study consists of 10 different machine learning classification algorithms, these algorithms are Logistic Regression, K-Nearest Neighbours, Support Vector Machine, Naive Bayes, Decision Tree, Random Forest, XGBoost, Gradient Boosting Machines (GBM), Light GBM, and Multi-Layer Perceptron Classifier (MLPC).

In a previous study [12], Support Vector Machine was stated to be the best of the algorithms of all. However, previous study’s comparison only includes model accuracy values. In imbalanced datasets, accuracy alone is insufficient since minority class has little effect on the metric [14]. Sections C and D inform about data imbalance and performance metrics. An enhanced comparison with multiple metrics is provided blow:

Table 1. Performance metrics of traditional machine learning algorithms

ML Algorithm	Model Accuracy	Precision 0	Precision 1	Recall 0	Recall 1	F1-Score 0	F1-Score 1
Logistic regression	0.49	0.54	0.40	0.61	0.33	0.58	0.36
KNN	0.50	0.51	0.50	0.64	0.37	0.57	0.43
SVM	0.63	1.0	0	0.64	0	0.78	0
Naive Bayes	0.54	0.49	0.65	0.71	0.42	0.58	0.51
Decision Tree	0.43	0.60	0.15	0.55	0.18	0.58	0.16
Random Forest	0.56	0.74	0.25	0.63	0.36	0.68	0.29
XGboost	0.62	0.69	0.50	0.71	0.48	0.70	0.49
GBM	0.56	0.69	0.35	0.65	0.39	0.67	0.37
Light GBM	0.51	0.63	0.30	0.61	0.32	0.62	0.31
MLPC	0.62	0.97	0	0.63	0	0.76	0

As seen in Table 1, F1-Score 1 indicates a tendency of classifying students as entrepreneurs. In Table 1, F1-Score 1 values show that machine learning algorithms are insufficient to predict entrepreneurs among students. High accuracy values of the algorithms stem from the high percentage of non-entrepreneur students in dataset.

2.3. Method

First of all, the dataset is balanced by applying handling imbalanced dataset algorithms. At this step, Random oversampling (ROS), Random Under-Sampling (RUS), SMOTE, and NearMiss methods are applied separately. Then, the results are obtained by applying machine learning methods to the balanced dataset. Finally, the results of the machine learning methods applied to the imbalanced dataset are compared with the machine learning methods applied to the dataset balanced using the handling imbalanced dataset techniques.

2.4. Imbalance data handling techniques

Imbalanced datasets are one of the main challenges for machine learning algorithms. A dataset is considered to be imbalanced if one of its classes plays a huge dominance over the rest of the classes [15]. Data imbalance is usually encountered with exception-based machine learning applications such as fraud detection, rare-disease identification, determining defective products and so forth.

Handling imbalanced dataset methods can be done in many ways. The most popular handle methods include Random oversampling (ROS), Random Under-Sampling (RUS), SMOTE, and NearMiss. Below are the explanations of different methods of imbalance data handling.

2.4.1. Random oversampling (ROS)

Random oversampling is a basic sampling method used for increasing the number of the minority class. Data points from the minor class are randomly selected and duplicated exactly in this method [16]. Resulting an increase, the number of minority samples to create a balance between both classes.

2.4.2. Random undersampling (RUS)

Random Undersampling is the simplest method of undersampling. Examples from the majority class are randomly selected and eliminated [16] to provide a balance between minority and majority class.

2.4.3. Synthetic minority oversampling technique (SMOTE)

Synthetic Minority Oversampling Technique (SMOTE) is based on creating new minority values around the original values. Minority class examples are linked with their neighbours and new synthetic values are created with their linear combination, forming all new values in the minority class area [17]. SMOTE method can be formulated as:

$$s = x + u \cdot (x' - x) \quad (1)$$

Here x is any sample, s is the new synthetic sample, x' is a randomly chosen value from the nearest neighbours of x and u is a random variable between 0 and 1 [18].

2.4.4. NearMiss

NearMiss is an under-sampling method. NearMiss algorithm clusters minority and majority class sample and then removes larger-class elements that are closest to smaller-class elements [19], increasing the differences between the two groups while decreasing the imbalance between the number of samples among two classes.

2.5. Performance Metrics

The key measure in the research concluded by Sharma and Manchanda [12] was accuracy. However, accuracy alone does not give a clear image of the whole picture [20]. There are other performance metrics as well for use of understanding if the algorithm has done well or not.

True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN) are four metrics in a confusion matrix. These metrics are used to define accuracy, recall, precision and F1 scores.

2.5.1. Accuracy

Accuracy is the number of correct guesses divided by all guesses.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

2.5.2. Precision

Precision is the correct positive guesses divided by total positive guesses.

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

2.5.3. Recall

Recall is the success rate of the guesses from all positive values.

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

2.5.4. F1-score

F1 score is the harmonic mean of recall and precision. It is used in imbalanced data especially because accuracy lacks in detecting False negatives. To put it better, in imbalanced data, %95 fraudulent transactions for instance; an algorithm that is literally identifying every action as a secure action would be %95 accurate. Because even though for every fraud the algorithm is wrong, there are so few of them to affect the whole accuracy compared to the total number of values.

F1 score, on the other hand, penalizes False-negative values. Since the F1 score is a harmonic mean, it has a tendency for the smaller value. Hence, a higher F1 score ensures a correct minority class prediction [14]. Therefore, even in imbalanced datasets, this score gives a clearer image.

$$F1\ Score = 2\frac{Precision\ Recall}{Precision + Recall} \quad (5)$$

3. Results

In this paper, the prediction of entrepreneurship competence with machine learning algorithms is studied as a result of tests on university students. In this context, base machine learning algorithms are used and an imbalance data is observed from the results. In order to eliminate this problem, the imbalanced data handling methods in the literature are used with machine learning algorithms. Base machine learning algorithms and machine learning algorithms with handling imbalanced data methods are compared with certain performance metrics. Successful results are obtained such as higher F1- Scores and model accuracy. It is seen that handling imbalanced data methods have a positive effect on the performance of machine learning algorithms and make noticeable contributions.

Undersampling methods such as RUS and NearMiss, have decreased the majority class samples to 91 to equate classes. Oversampling methods such as ROS and SMOTE, have increased the minority class samples to 128 to equate classes.

Using handling imbalanced data methods have shown to have better F1-Scores on all algorithms. Machine learning algorithms without using imbalance data handling techniques have performed significantly poorly on F1 Scores for entrepreneurs, implying a failure considering algorithms were to predict entrepreneurial competency. Every imbalanced data handling method has increased the F1 Score-1 performance of machine learning algorithms.

The best results are acquired by Random Forest algorithm with using RUS method. In this case, not only accuracy of the model increases by %35 to 0.76, but also high F1 scores for both output 0 and 1 are attained, indicating higher precision and recall scores as well.

The Logistic Regression algorithm has been observed to increase model accuracy for every imbalanced data handling method with the best results obtained by using SMOTE. Overall model accuracy is increased by 28% to 0.63. Performance metrics of every algorithm based on model accuracy, precision, recall and F1-Scores are tabulated below. Precision values, recall values, and F1 Scores are all improved. A 0.59 F1-Score 1 is attained. The results about logistic regression are shown in Table 2.

Table 2. Logistic regression with imbalanced data handling methods

ML Algorithm	Model Accuracy	Precision 0	Precision 1	Recall 0	Recall 1	F1-Score 0	F1-Score 1
RUS-LR	0.59	0.44	0.76	0.69	0.53	0.54	0.63
ROS-LR	0.63	0.67	0.57	0.67	0.57	0.67	0.57
NM-LR	0.57	0.60	0.52	0.60	0.52	0.60	0.52
SMOTE-LR	**0.63**	**0.64**	**0.61**	**0.68**	**0.57**	**0.66**	**0.59**

The model accuracy of K-Nearest Neighbours is increased by %18 to 0.59 with SMOTE method. A slight increase and improvement in the balance between F1 scores are also achieved. The results about KNN algorithm are shown in Table 3.

Table 3. K-Nearest neighbours with imbalanced data handling methods

ML Algorithm	Model Accuracy	Precision 0	Precision 1	Recall 0	Recall 1	F1-Score 0	F1-Score 1
RUS-KNN	0.54	0.56	0.52	0.58	0.50	0.57	0.51
ROS-KNN	0.58	0.47	0.71	0.68	0.51	0.56	0.60
NM-KNN	0.43	0.36	0.52	0.47	0.41	0.41	0.46
SMOTE-KNN	**0.59**	**0.42**	**0.82**	**0.75**	**0.52**	**0.54**	**0.64**

Support Vector Machine algorithm is observed to lose overall model accuracy by %17.4 down to 0.52. However, without implementing imbalanced data handling methods SVM is unable to detect *any entrepreneurs.* The algorithms without implementing imbalanced data handling methods and a model accuracy of 63% was obtained in SVM. Nevertheless, recall 1, precision 1 and F1 Score 1 are 0 in SVM, indicating that this algorithm fails at classifying output 1 (entrepreneurs). Hence, a correct interpretation for this algorithm would mark that after handling imbalanced data the results have been improved. The results about SVM algorithm are shown in Table 4.

Table 4. SVM with imbalanced data handling methods

ML Algorithm	Model Accuracy	Precision 0	Precision 1	Recall 0	Recall 1	F1-Score 0	F1-Score 1
RUS-SVM	**0.52**	**0.16**	**0.95**	**0.80**	**0.49**	**0.27**	**0.65**
ROS-SVM	0.41	0	0.93	0	0.42	0	0.58
NM-SVM	0.47	0.20	0.81	0.56	0.46	0.29	0.59
SMOTE-SVM	0.44	0	1	0	0.44	0	0.61

Naive Bayes algorithm demonstrated performance improvements for each imbalanced data handling method. The most improved method is SMOTE with an overall model accuracy of 0.67. Also, every method has elevated both F1 Scores. The results about Naive Bayes algorithm are shown in Table 5.

Imbalanced Data Handling methods have amended Decision Tree algorithm as well, with the optimal solution Decision Tree with RUS applied. It is observed that every F1 Score 1 is escalated along with model accuracies. On

the other hand, all F1 Score 0 values except the application of the NearMiss method have also increased. The results about Decision Tree algorithm are shown in Table 6.

Table 5. Naive Bayes with imbalanced data handling methods

ML Algorithm	Model Accuracy	Precision 0	Precision 1	Recall 0	Recall 1	F1-Score 0	F1-Score 1
RUS-NB	0.63	0.64	0.62	0.67	0.59	0.65	0.60
ROS-NB	0.63	0.61	0.64	0.69	0.56	0.65	0.60
NM-NB	0.59	0.60	0.57	0.62	0.55	0.61	0.56
SMOTE-NB	**0.67**	**0.67**	**0.68**	**0.73**	**0.61**	**0.70**	**0.64**

Table 6. Decision tree with imbalanced data handling methods

ML Algorithm	Model Accuracy	Precision 0	Precision 1	Recall 0	Recall 1	F1-Score 0	F1-Score 1
RUS-DT	**0.74**	**0.80**	**0.67**	**0.74**	**0.74**	**0.77**	**0.70**
ROS-DT	0.63	0.64	0.68	0.72	0.59	0.68	0.63
NM-DT	0.54	0.52	0.57	0.59	0.50	0.55	0.53
SMOTE-DT	0.64	0.61	0.68	0.71	0.58	0.66	0.62

Random Forest algorithm provided best prediction performance among all algorithms, as stated before. The optimal method to use with this algorithm is founded to be RUS. Overall model accuracy has been increased from 0.56 to 0.76. F1 Score 1 feature is significantly ameliorated to %70 success alongside a 17% increase in F1 Score 0 to 80% success. The results about Random Forest algorithm are shown in Table 7.

Table 7. Random forest with imbalanced data handling methods

ML Algorithm	Model Accuracy	Precision 0	Precision 1	Recall 0	Recall 1	F1-Score 0	F1-Score 1
RUS-RF	**0.76**	**0.88**	**0.62**	**0.73**	**0.81**	**0.80**	**0.70**
ROS-RF	0.74	0.67	0.82	0.83	0.66	0.74	0.73
NM-RF	0.46	0.48	0.43	0.50	0.41	0.49	0.42
SMOTE-RF	0.72	0.64	0.82	0.82	0.64	0.72	0.72

XGB method acquired an accuracy improvement with two methods, RUS and ROS to be precise. Although SMOTE and NearMiss seem to have lowered the overall accuracy, SMOTE method has refined F1 Score 1 from 0.49 to 0.58. The results about XGB algorithm are shown in Table 8.

Table 8. XGboost with imbalanced data handling methods

ML Algorithm	Model Accuracy	Precision 0	Precision 1	Recall 0	Recall 1	F1-Score 0	F1-Score 1
RUS-XGB	**0.72**	**0.80**	**0.62**	**0.71**	**0.72**	**0.75**	**0.67**
ROS- XGB	0.64	0.58	0.71	0.72	0.57	0.65	0.63
NM-XGB	0.50	0.48	0.52	0.55	0.46	0.51	0.49
SMOTE-XGB	0.61	0.61	0.61	0.67	0.55	0.64	0.58

Gradient Boosting Machine (GBM) algorithm performed %28 better with the application of SMOTE method. Both F1 Scores have increased. Implying the effectiveness of imbalanced data handling. The results about GBM algorithm are shown in Table 9.

Table 9. GBM with imbalanced data handling methods

ML Algorithm	Model Accuracy	Precision 0	Precision 1	Recall 0	Recall 1	F1-Score 0	F1-Score 1
RUS-GBM	0.67	0.84	0.48	0.66	0.71	0.74	0.57
ROS- GBM	0.69	0.72	0.64	0.72	0.64	0.72	0.64
NM- GBM	0.52	0.48	0.57	0.57	0.48	0.52	0.52
SMOTE-GBM	**0.72**	**0.67**	**0.79**	**0.80**	**0.65**	**0.73**	**0.81**

Light GBM algorithm is also seen to improve in every metric. In the optimal case, RUS increased this algorithm's modal accuracy to 0.72. Furthermore, F1 Score 0 and 1 have attained 0.75 and 0.67 successes respectively. The results about Light GBM algorithm are shown in Table 10.

Table 10. LGBM with imbalanced data handling methods

ML Algorithm	Model Accuracy	Precision 0	Precision 1	Recall 0	Recall 1	F1-Score 0	F1-Score 1
RUS-LGBM	**0.72**	**0.80**	**0.62**	**0.71**	**0.72**	**0.75**	**0.67**
ROS- LGBM	0.69	0.67	0.71	0.75	0.62	0.71	0.67
NM- LGBM	0.65	0.64	0.67	0.70	0.61	0.67	0.64
SMOTE-LGBM	0.62	0.61	0.64	0.69	0.56	0.65	0.60

Model accuracy of MLPC algorithm seems only improved with ROS method. However, as seen in table 1 MLPC algorithm without imbalanced data handling methods has failed in precision 1, recall 1, and F1 Score 1 values. On the contrary, all imbalanced data handling methods have increased the performance metrics of entrepreneurship prediction. These statistics indicate that not only ROS but every imbalanced data handling method used in this paper has improved the algorithm from a certain perspective. The results about MLPC algorithm are shown in Table 11.

Table 11. MLPC with imbalanced data handling methods

ML Algorithm	Model Accuracy	Precision 0	Precision 1	Recall 0	Recall 1	F1-Score 0	F1-Score 1
RUS-MLPC	0.59	0.96	0.14	0.57	0.75	0.72	0.24
ROS- MLPC	**0.63**	**0.61**	**0.64**	**0.69**	**0.56**	**0.65**	**0.60**
NM- MLPC	0.61	0.68	0.52	0.63	0.58	0.65	0.55
SMOTE-MLPC	0.58	0.61	0.54	0.63	0.52	0.62	0.53

4. Discussion and Conclusion

Analysis mainly focuses on the effect of different imbalance data handling techniques to improve the comparison metrics which are accuracy, precision, recall and F1-score to predict the entrepreneurial competency in university students. Python programming language was used to perform the analysis.

This analysis shows that imbalance data handling techniques have a significant effect in machine learning. Previous research conducted on the same dataset provided decent solutions proposals. The SVM algorithm was found to be the best algorithm in previous work. However, a broad perspective of performance metrics proved that the results could be improved. Machine learning algorithms without imbalanced data handling methods lacked detecting entrepreneurial competency. As expected, Imbalanced Data handling methods increased the F1 Scores, in this case, entrepreneurial competency as an output of 0 or 1. This research does not only imply the effects of imbalanced data handling methods on this particular topic but also many machine learning algorithms that can be carried out later. Imbalanced data handling methods have been used alongside machine learning algorithms, nevertheless, these methods have broader application areas.

The results may provide an insight into a person, which can help people improve their entrepreneurial skills. These results can also be localized. This work may be used in future software of an HR company that studies larger groups of people, eventually localizing for countries or even sectors.

In this paper, we used the same dataset in Sharma [12]'s work was used. Sharma [12] compared the base Machine Learning algorithms Random Forest, K-Nearest Neighbours, Support Vector Machine (SVM), Logistic Regression, Naive Bayes and Decision Tree algorithms and the best algorithm was found to be SVM with a model accuracy of 59.18%. However, Sharma [12] did not use other performance measures such as F1 Score 1, precision 1, and recall 1. In this paper, higher model success was obtained than the previous study using imbalanced data handling methods.

In this paper, the effects of imbalanced data handling methods on machine learning algorithms have been examined. The research is conducted on a dataset of 219 Indian university students. Traditional machine learning algorithms were trained by 75% of the data and tested by 25%. The results without implementing imbalanced data handling methods lacked detecting entrepreneurs among others. Even though most algorithms have similar accuracy values the reason was the insufficient number of entrepreneurs. Applying imbalance data handling techniques to data with base machine learning algorithms provided satisfactory results. Higher accuracy values and F1 Scores are obtained with these techniques. The best results were found with the Random Forest algorithm using the RUS method with a 0.76 model accuracy and 0.80 F1 Score 1. It is obvious from comparing the algorithm without using the methods, that the algorithm has upgraded its prediction performance from 0.29 to 0.80.

References

[1] R. C. Ramona, "The Importance Of Entrepreneurs In The 'New Economy'", Managerial Challenges of the Contemporary Society, Issue 2, pp. 265-269, Jun. 2011.

[2] R. W. Fairlie and W. Holleran, "Entrepreneurship training, risk aversion and other personality traits: Evidence from a random experiment", Journal of Economic Psychology, vol. 33, pp. 366-378, Apr. 2012.

[3] S. P. Kerr, W. R. Kerr and M. Dalton, "Risk attitudes and personality traits of entrepreneurs and venture team members", Proceedings of the National Academy of Sciences of the United States of America, vol. 116, pp. 17712-17716, Sep. 2019.

[4] M. Brandt and S. STEFÁNSSON, "The personality venture capitalists look for in an entrepreneur: An artificial intelligence approach to personality analysis", M. Sci. thesis, KTH Royal Institute of Technology, Stockholm, Sweden, Jun. 2018.

[5] M. Caliendo, F. Fossen and A. S. Kritikos, "Personality characteristics and the decisions to become and stay self-employed", Small Business Economics, vol. 42, pp. 787-814, Oct. 2013.

[6] F. U. Salmony and D. K. Kanbach, "Personality trait differences across types of entrepreneurs: a systematic literature review", Review of Managerial Science, vol. 16, pp. 713-749, Apr. 2021.

[7] M. Castillo-Palacio, R. M. Batista-Canino, A. Zuñiga-Collazos, "The Relationship between Culture and Entrepreneurship: From Cultural Dimensions of GLOBE Project", Scientific Annals Of Economics and Business, vol. 67, pp. 517-532, Mar. 2020.

[8] C. J. Boudreaux, B. N. Nikolaev, and P. Klein, "Socio-cognitive traits and entrepreneurship: The moderating role of economic institutions", Journal of Business Venturing, vol. 34, pp.178-196, Jan. 2019.

[9] D. V. Moudrý and P. Thaichon, "Enrichment for retail businesses: How female entrepreneurs and masculine traits enhance business success", Journal of Retailing and Consumer Services, vol. 54, May 2020.

[10] B. Graham and K. Bonner, "One size fits all? Using machine learning to study heterogeneity and dominance in the determinants of early-stage entrepreneurship", Journal of Business Research, vol. 152, pp.42-59, Nov. 2022.

[11] M. G. Celbiş, "A machine learning approach to rural entrepreneurship", Papers in Regional Science, vol. 100, pp. 1079-1104, Jan. 2021.

[12] U. Sharma and N. Manchanda, "Predicting and Improving Entrepreneurial Competency in University Students using Machine Learning Algorithms", in 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 2020, pp. 305-309.

[13] N. Manchanda and U. Sharma (2019) [Online]. Available: https://www.kaggle.com/datasets/namanmanchanda/entrepreneurial-competency-in-university-students

[14] Y. Sun, A. K. C. Wong, and M. S. Kamel, "Classification of Imbalanced Data: A Review", International Journal of Pattern Recognition and Artificial Intelligence, vol. 23, pp. 687-719, 2009.

[15] A. Somasundaram and U. S. Reddy, "Data Imbalance: Effects and Solutions for Classification of Large and Highly Imbalanced Data", in 1st International Conference on Research in Engineering, Computers and Technology (ICRECT 2016), 2016.

[16] H. Ali et al., "A review on data preprocessing methods for class imbalance problem", International Journal of Engineering &Technology, vol. 8, pp. 390-397, 2019.

[17] N. V. Chawla et al., "SMOTE: Synthetic Minority Over-sampling Technique", Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, Jun. 2002

[18] R. Blagus and L. Lusa, "SMOTE for high-dimensional class-imbalanced data", in 11th International Conference on Machine Learning and Applications Machine Learning and Applications, 2012, paper 2, pp. 89-94.

[19] Md. A. Sahid et al., "Effect of Imbalance Data Handling Techniques to Improve the Accuracy of Heart Disease Prediction using Machine Learning and Deep Learning", in IEEE Region 10 Symposium (TENSYMP), 2022.

[20] M.V. Joshi, V. Kumar, and R.C. Agarwal, "Evaluating Boosting Algorithms to Classify Rare Classes: Comparison and Improvements", in Proceedings 2001 IEEE International Conference on Data Mining, 2001, pp. 257-264.