I.J. Information Engineering and Electronic Business, 2021, 6, 13-21
Published Online December 2021 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijieeb.2021.06.02

An Efficient Smote-based Model for Dyslexia Prediction

Vani Chakraborty
Research Scholar, Garden City University

Meenatchi Sundaram
Associate Professor, Garden City University

Received: 05 August 2021; Accepted: 23 September 2021; Published: 08 December 2021

Abstract: Dyslexia is a learning disability which causes difficulty in an individual to read, write and spell and do simple mathematical calculations. It affects almost 10% of the global population and detecting it early is paramount for its effective handling. There are many different methods to detect the risk of Dyslexia. Some of these methods are using assessment tools, handwriting recognition, expert psychological help and also using the eye movement data recorded while reading. One of the other convenient and easy ways of detecting risk of dyslexia is to make an individual participate in a simple game related to phonological awareness, syllabic awareness, auditory discrimination, lexical awareness, visual working memory, and many more and recording the observations. The proposed research work presents an effective way of predicing the risk of dyslexia with high accuracy and reliability. It uses a dataset made available from the kaggle repository to predict the risk of dyslexia using various machine learning algorithms. Also it is observed that the dataset has an unequal distribution of positive and negative cases and so the classification accuracy is compromised if used directly. The proposed research work uses three resampling techniques to reduce the imbalance in the dataset. The resampling techniques used are undersampling using near-miss algorithm, oversampling using SMOTE and ADASYN. After applying the undersampling near-miss algorithm, best accuracy was given by SVC classifier with the value of 81.63%. All the other classifiers used in the experiment produced accuracy in the range of 64% to 79.08%. After using the oversampling algorithm SMOTE, the classifiers produced very good results in the evaluation metrics of accuracy,CV score, F1 Score and recall. The maximum accuracy was given by RandomForest with a value of 96.37% and closely followed by XGBBoosting and GradientBoosting with an accuracy of 95.14%. Decision tree, SVC and ADABoost got an accuracy of 91.26%, 93.36% and 93.48% respectively. Even the values of CV score, F1 and recall were considerably high for all these classifiers. After applying the oversampling technique of ADASYN, RandomForest algorithm generated maximum accuracy of 96.25%. Between the two oversampling techniques, SMOTE algorithm performed slightly better in producing better evaluation metrics than ADASYN. The proposed system has very high reliability and so can be effectively used for detecting the risk of dyslexia.

Index Terms: Dyslexia, gamified test, SMOTE, ADASYN, Near-miss.

1. Introduction

Specific learning disability refers to a varied group of conditions where an individual experiences difficulty in processing language, be it spoken or written. This condition manifests as a difficulty to speak, read, write, spell and do simple mathematical calculations and each of these disabilities are referred as dyslexia, dysgraphia, dyscalculia and dyspraxia. Dyslexia also called as the reading disability is one of the most common learning disabilities and account for atleast 80% of all learning disabilities[1]. According to the International Dyslexia Association, Dyslexia is a neurological disorder. There is no specific cure for the condition and if detected early will be very helpful to design the coping strategies and learning methods[2]. There are many methods available for detecting dyslexia. One of the most important methods is to measure the behavioural symptoms particularly noticeable in individuals having dyslexia. These symptoms are related to reading, writing and phonological awareness. With the help of expert psychologists conducting standardised tests such as WISC, WJ, CELF, OWLS[4] and many more. But conducting such tests are very time consuming, expensive and very difficult to implement in large populations[3,4]. Some of the other methods of detecting dyslexia are studying brain images, recording the eye movement of an individual while reading and predicting the risk of dyslexia, handwriting recognition and collecting questionnaire[5,6]. In order to make the process of prediction much simpler and easier, an online game based test is designed and the data collected from such a test is put

through various machine learning models to predict the risk of dyslexia[7]. The important purpose of this research work is to generate an efficient model for predicting the risk of Dyslexia in individuals. The dataset used in this work is collected from responses in an online linguistic game, which can be collected in a much easier way compared to other methods. In this paper, the data collected from online gamified test is used to analyse and predict the risk of dyslexia. The dataset used for this is available for download from kaggle[7,16].

2. Literature Review

Dyslexia is one of the common learning disability affecting 10 to 15% of the world's population. It affects 10 to 17% of the population in the USA and 8.6% to 11% in Spain[8]. Various machine learning models and optimization techniques are developed to predict the risk of dyslexia in individuals.

Prediction of learning disabilities in school-age children is achieved by two machine learning approaches, rough sets and decision trees. In rough sets, attribute reduction was achieved using Johnson's reduction algorithm and classification was achieved using Naïve Bayes algorithm. For the construction of decision trees, J48 algorithm was used. It was observed that the performance of decision tree was poor in various aspects compared to rough sets.The data that has been used in this study was collected through a questionnaire from the stakeholders including the parent, teachers and the individual. The responses are collected as binary variables having a yes/no answer. A total of 14 features were collected and used for this study[9].

Dyslexia is a neurological disorder and the cause for it is still not clear. Many individuals face a clear difficulty in reading and spelling. It is easy to detect the difficulty experienced by a young reader by tracking the eye movement measures when they are reading a textual content. Using various predictive modelling techniques, classification models were developd that can detect the risk of dyslexia. Eye movement in reading can be used to predict long term reading difficulties in children. The best accuracy was obtained by the SVM-RFE classifier with an accuracy of 95.3%+-4.5%[10]. In this study, the dataset used has been the eye tracking data when an individual is reading a paragraph of text. Suitable algorithms are used for identifying the features from the raw eye tracking data[10].

Features can be extracted from the brain electrical signals using the EEG scan dataset. For the early prediction of dysleia, a computational classifier was developed using k-Means, ANN and Fuzzy logic algorithm. An accuracy of 89.6% was obtained for k-Means. ANN and Fuzzy logic gave an accuracy of 89.7% and 85.7% respectively[5]. In this study brain electrical signals collected using EEG scan were used as dataset[5].

To differentiate dyslexic from non-dyslexics eye movement data was used to develop a predictive model. Particle Swarm Optimization technique was used to optimize the prediction accuracy. Principal Component Analysis was used for feature extraction. Linear SVM produced an accuracy of 90% and SVM-PSO produced an accuracy of 95%[11]. This research work also made use of the eye tracking data with features extracted and the features were used for predicting the risk of dyslexia[11].

Various other research works have used different kinds of data for predicting the risk of dyslexia. Some of these methods used for data collection are not simple and straight forward. To make the detection of dyslexia easier, an online gamified test along with a predictive machine learning model was used. Administering this online game test is also much easier compared to data collection used in other methods. More than 3600 participants were involved in this study and the model predicted the risk of dyslexia with an accuracy over 80%. The robustness of the model was checked with another dataset of 1300 participants having participated in the gamified test. Random forest algorithm was used to develop the predictive model. It was observed that best results were obtained for the age category of 9-11 and lot of emphasis was given on the sensitivity analysis of classifiers[7].

3. Dataset

The dataset that is used in the proposed research work is "Predicting risk of Dyslexia" from the Kaggle repository. This study was approved by Carnegie Mellon University Institutional Review Board and was conducted by the author Luz Rello et al . The purpose of this study was to understand whether the risk of dyslexia can be predicted or not, by administering questions through an online game, that is very easy to implement[7]. Hence this dataset was considered for the proposed work. A total of 3644 participants were involved in this study in the age group of 7 – 17 years. Of this 392 were dyslexics and the rest were control population. All the tests were conducted in the Spanish language. Four features related to the demography were collected from the participants. They are Gender, NativeLang(to understand whether their native lang is Spanish or not), LangSubject(a binary value of yes or no to understand whether the participant has failed in any language subject in school) and age. Apart from the 4 demography related features, answers related to 32 questions were collected from the participants through games. Question nos 1 – 21 had questions related to auditory and visual discrimination. Question no 22-29 were focussed on correcting words and sentences. Question no 30-32 were used for checking sequential visual and auditory working memory. For each of the questions 1 – 32, 6 responses were recorded. They are NoClicks, number of correct answers(hits), sum of hits(score), accuracy which is calculated as number of hits/number of clicks, missrate which is calculated as no of misses / no of clicks. So a total of 4

demography related features + 32 questions X 6 responses for each question = 192. The total number of features was 196. Since out of 3644 participants, only 392 were dyslexics and so the dataset is clearly imbalanced[7]. There were no missing values in the dataset. Output variable is a column called 'Dyslexia' which has a categorical value of Yes or No indicating the presence or absense of dyslexia. The following table describes the dataset and its features[7,16].

Table 1. Dataset Description

S.no	Data Type	Attribute	Description
1	Categorical	Gender	Gender of the participant (Values – Male,Female)
2.	Categorical	Nativelang	Yes or No indicating whether the indivdual's native language is Spanish
3.	Categorical	Otherlang	Yes or No indicating whether the individual has failed in any language subject so far in academics
4.	Numeric	Age	Described in years(7 – 17 years old)
5.	Numerica	Clicks1	Number of clicks
6	Numeric	Hits1	Number of correct answers
7.	Numeric	Misses1	Number of incorrect answers
8.	Numeric	Score1	Sum of hits
9.	Numeric	Accuracy1	Number of hits / Number of clicks
10.	Numeric	Missrate1	Number of misses / Number of clicks
Features No 5 to 10 are repeated for 31 more questions of the game			
11.	Categorical	Dyslexia	Yes or No indicating whether the individual has dyslexia or not

The first 4 features of the dataset are demographic features and from 5 till 196 are related to the answers given by the participant in the game. For each of the question, 6 important information were collected and they are NoClicks, hits, misses, score, accuracy and missrate. So, for 32 questions there are 32 X 6 = 192 features that were collected. Adding the 4 demographic features to the same and a total of 196 features are collected for each participant of the online game. There is also a target variable called "Dyslexia" which is either a Yes or No indicating whether the participant has Dyslexia or not. A graphical study of the dataset is given below.

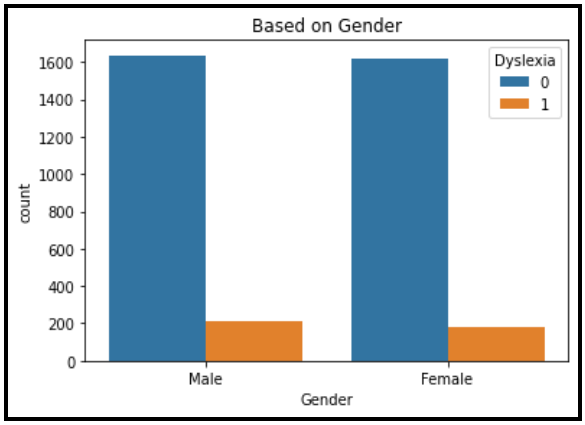


Fig.1. Target distribution based on gender

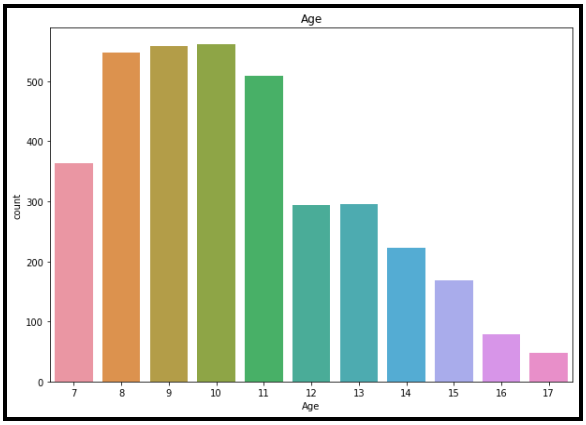


Fig.2. Age group of the population studied

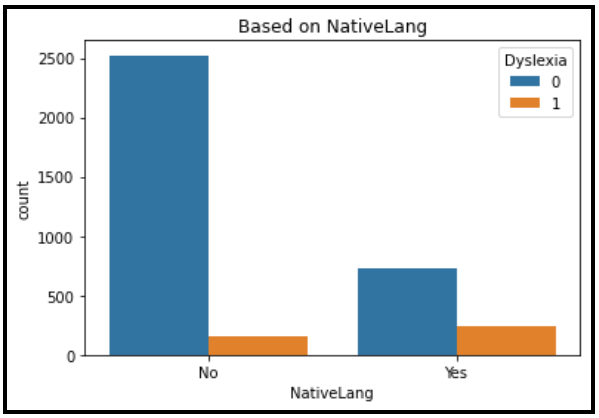


Fig 3. Target distribution based on NativeLang

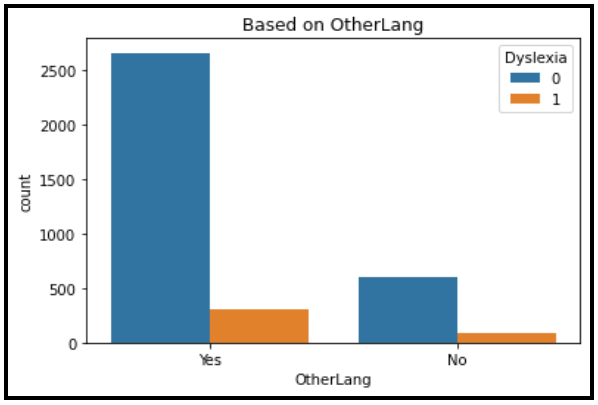


Fig. 4. Target distribution based on failure in lang subject

Fig.1 displays the target distribution based on gender. Fig. 2 shows the distribution in the population age group of the participants. From the graph of Fig.2 it can be observed that the age group of the participants was from 7 to 17 and the maximum participants were from the age group of 11. In order to understand whether the participants of the study had their NativeLang as Spanish, a binary response was collected and a graph generated to see the target distribution in NativeLang. Also in order to observe whether the participants have failed, in any language related subject so far in their academic pursuit, one more feature was collected and a graph depicting their response with respect to the target distribution is displayed in Fig.4.

4. Imbalance Nature of Dataset

The dataset contains 3252 instances of the negative class [No/0] and 392 instances of the positive class [Yes/1]. Clearly there is an unequal distribution of the classess in the dataset. Since the dataset is skewed, if a classifier is used on this dataset for prediction, there would be bias in the decision making process. Since the non dyslexics are way more than the dyslexia affected individual, the prediction would be skewed towards the majority class[20]. The classification models are not able to make accurate prediction because of this unequal distribution.

The problem of the classification inaccuracy on imbalanced dataset can be handled at two levels and they are at the data level and algorithm level. At the data level, oversampling and undersampling methods are used to balance the data for classification reconstruction[15]. One of the important contributions of the proposed work is to understand the techniques for handling imbalanced data. One of the important approach for handling classification of imbalanced data is the resampling techniques and apart from that, cost-sensitive methods and ensemble methods are also used for the same[12]. In the resampling technique, there are three important methods and they are undersampling majority class, and oversampling minority class by applying SMOTE (Synthetic Minority Oversampling Technique) and ADASYN (Adaptive Synthetic sampling approach). In this proposed work undersampling of majority class is implemented using Near-miss algorithm and oversampling of minority class is achieved using SMOTE and ADASYN to understand the effective technique on the dataset considered.

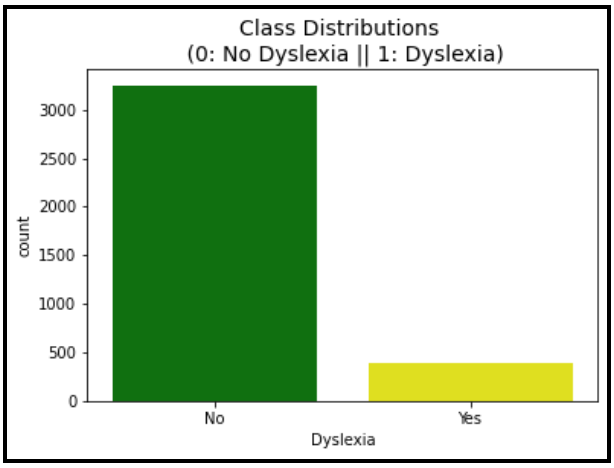


Fig. 5. Imbalance in the class distribution

5. Dataset Resampling Techniques

In statistical terms, sampling is a subset of the selected population. The purpose of sampling is to ensure that the sampling group which is chosen is a true representation of the population [13]. Even though the dataset is chosen after careful sampling, it might have to be resampled to improve it further for better results. Resampling is performed to adjust the class distribution when particularly dealing with imbalanced dataset [14]. Undersampling methods resample the data in such a way that they reduce some samples from the majority class. It is achieved by randomly removing samples (random sampling), or by using statistical information (informed under sampling) [15]. A major drawback of undersampling is that we could remove some information which is used for classifying the data. Oversampling methods add samples to the minority class in order to make it similar to the majority class. The new samples that are added can be original data, randomly selected from the data source or synthetic data that are added by certain algorithms like SMOTE or ADASYN [15]. SMOTE stands for Synthetic Minority Over Sampling Techniques.SMOTE preprocessing technique is considered to be one of the most powerful and reliable in the field fo machine learning for working with imbalanced dataset[21]. It is a very popular approach for the construction of a classifier for the imbalanced dataset. Imblearn library is used for the implementation of SMOTE. SMOTE technique is applied to the given dataset to generate equal number of positive and negative values. ADASYN stands for Adaptive Synthetic Sampling approach. It works similar to SMOTE in that this approach also generates synthetic observations for the minority class. But here synthetic observations are generated for those that are harder to learn than those that are easier to learn. A comparative study of the performance of ADASYN and SMOTE is done to observe the prediction accuracy result. Various machine learning algorithms are applied to the generated dataset and the results recorded. Generally the performance of a classificication model is evaluated by looking at its F1-score, precision, accuracy and recall. But when we consider the real world data it is important to understand the cross validation score so that overfitting of the data can be avoided. Here a cross validation of 5 folds are performed on the classification models and their cv scores are recorded and compared. Also F1-score and recall are also observed to understand how the different classifiers perform.

6. Proposed framework and Experimental Results

The proposed framework is explained in the flowchart given in Figure 6. The Dyslexia prediction dataset does not have any null values and so no preprocessing is required. In the next step, undersampling near-miss algorithm is applied on the dataset to generate equal number of positive and negative cases and to balance the dataset.

Also, on the dataset, SMOTE oversampling technique and ADASYN oversampling technique are applied. At the final stage, various machine learning algorithms are applied on the given dataset and results obtained. Also a k-fold validation with a validation of 5 was applied on the dataset to obtain better results. In order to compare the efficiency of the different sampling technique, the evaluation metrics of accuracy, cross validation score, F1 score and recall score are studied.

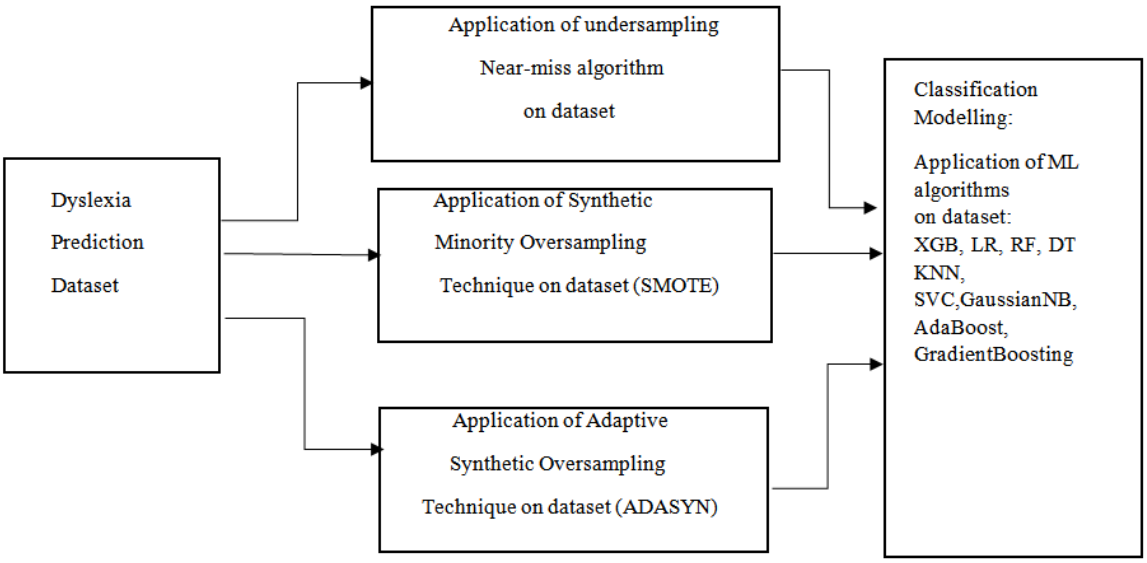
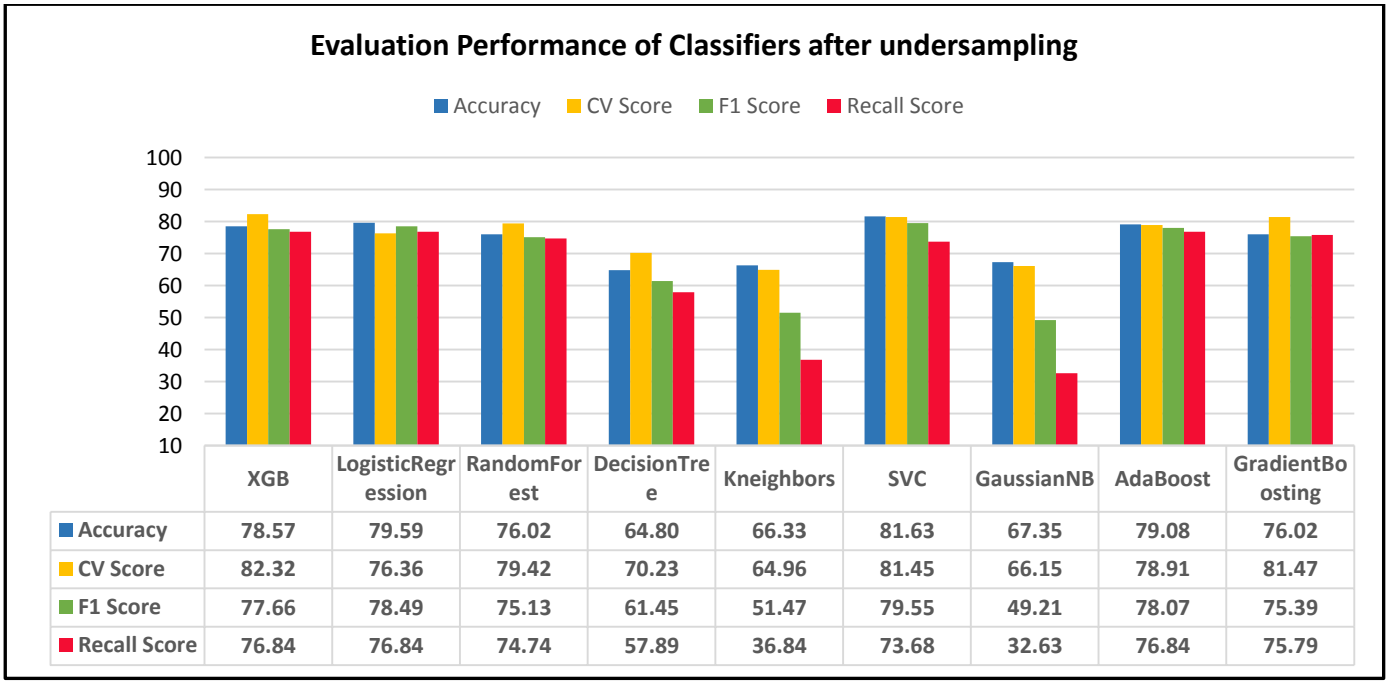


Fig. 6. Proposed Framework

A. Undersampling using Near-miss algorithm

Near-miss is an algorithm that is useful for balancing an imbalanced dataset. It is an efficient way to balance the data. The algorithm does this by understanding the class distribution and randomly eliminates samples from larger classes. Near-miss algorithm finds the data with the least average distance to the negative class's nearest sample [19].

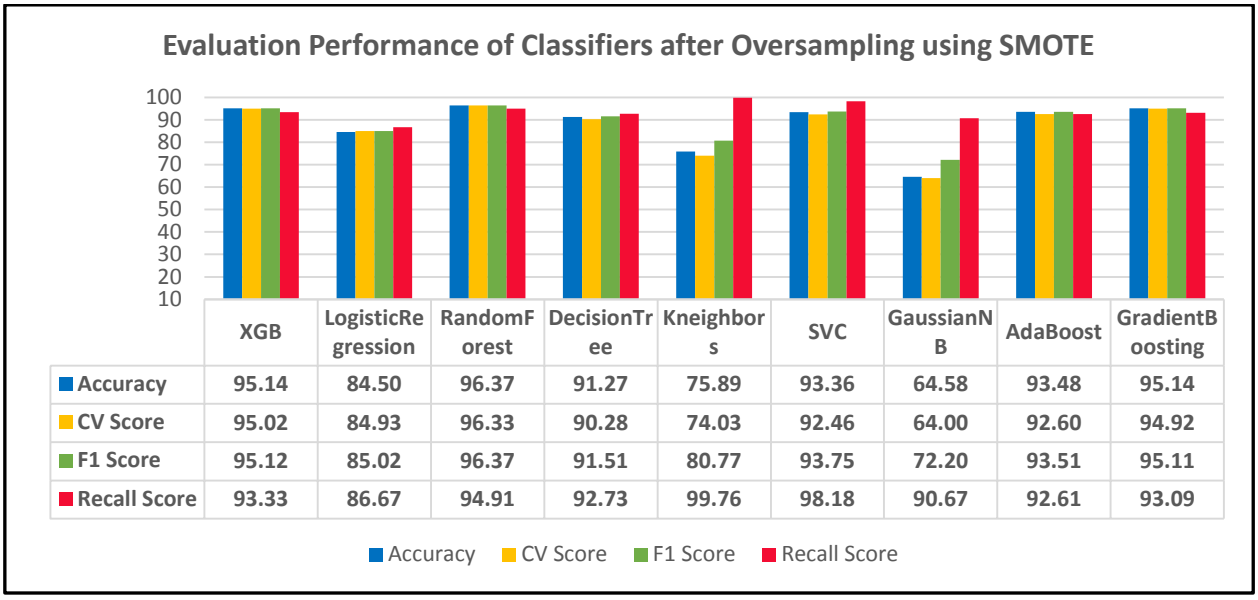


	XGB	LogisticRegression	RandomForest	DecisionTree	Kneighbors	SVC	GaussianNB	AdaBoost	GradientBosting
Accuracy	78.57	79.59	76.02	64.80	66.33	81.63	67.35	79.08	76.02
CV Score	82.32	76.36	79.42	70.23	64.96	81.45	66.15	78.91	81.47
F1 Score	77.66	78.49	75.13	61.45	51.47	79.55	49.21	78.07	75.39
Recall Score	76.84	76.84	74.74	57.89	36.84	73.68	32.63	76.84	75.79

Fig.7. Evaluation Performance of classifiers after undersampling

B. Oversampling using SMOTE

SMOTE stands for Synthetic Minority over Sampling Techniques. It is a very popular approach for the construction of a classifier for the imbalanced dataset. Imblearn library is used for the implementation of SMOTE. In the SMOTE technique, new instances are synthesized to balance the dataset. SMOTE technique is applied to the given dataset to generate equal number of positive and negative values[17]. For every underrepresented instance, first a predetermined number of neighbours are calculated and then some minority class instances are randomly chosen for creating the synthetic data[17].

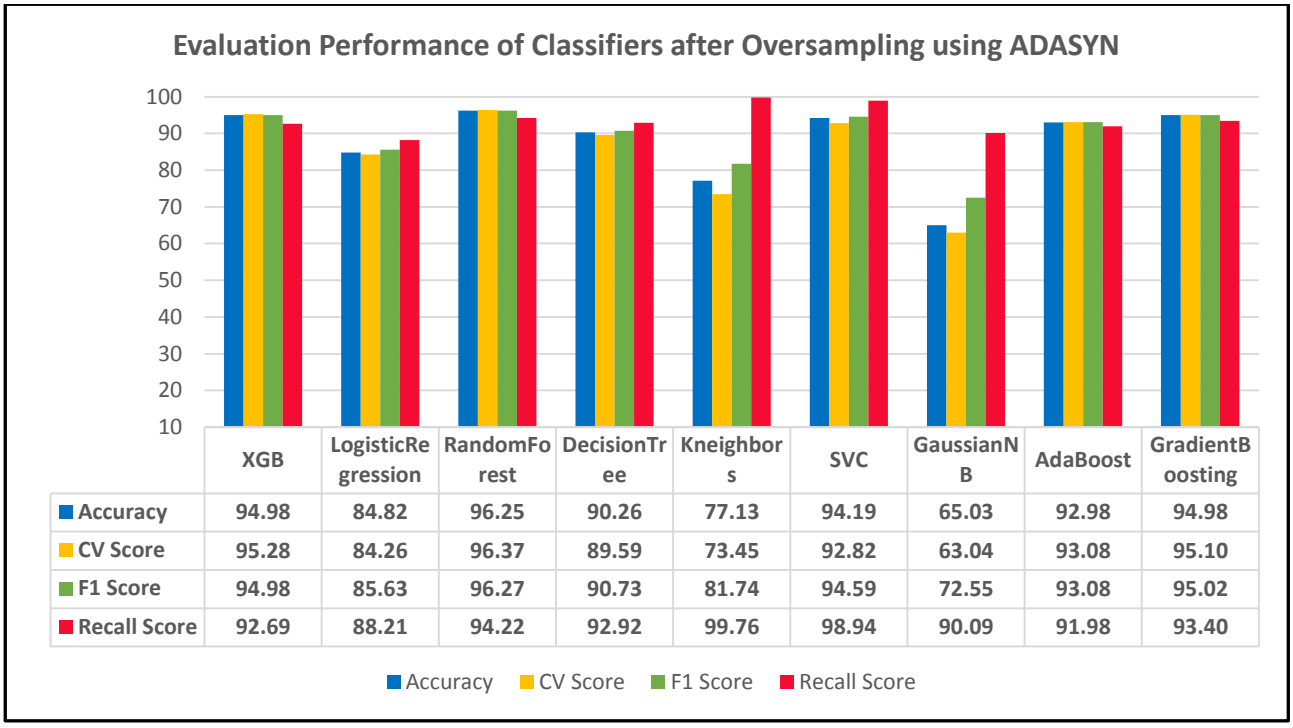


	XGB	LogisticRegression	RandomForest	DecisionTree	Kneighbors	SVC	GaussianNB	AdaBoost	GradientBoosting
Accuracy	95.14	84.50	96.37	91.27	75.89	93.36	64.58	93.48	95.14
CV Score	95.02	84.93	96.33	90.28	74.03	92.46	64.00	92.60	94.92
F1 Score	95.12	85.02	96.37	91.51	80.77	93.75	72.20	93.51	95.11
Recall Score	93.33	86.67	94.91	92.73	99.76	98.18	90.67	92.61	93.09

Fig.8 . Evaluation Performance of Classifiers after oversampling using SMOTE

C. Oversampling using ADASYN

ADASYN stands for Adaptive Synthetic Sampling approach. It works similar to SMOTE in that this approach also generates synthetic observations for the minority class. But here synthetic observations are generated for those that are harder to learn than those that are easier to learn. In this method certain weights are assigned to the minority class depending on their difficulty of being learned. ADASYN generates more synthetic data for the more difficult samples than the easier ones[18].



	XGB	LogisticRegression	RandomForest	DecisionTree	Kneighbors	SVC	GaussianNB	AdaBoost	GradientBoosting
Accuracy	94.98	84.82	96.25	90.26	77.13	94.19	65.03	92.98	94.98
CV Score	95.28	84.26	96.37	89.59	73.45	92.82	63.04	93.08	95.10
F1 Score	94.98	85.63	96.27	90.73	81.74	94.59	72.55	93.08	95.02
Recall Score	92.69	88.21	94.22	92.92	99.76	98.94	90.09	91.98	93.40

Fig.9. Evaluation Performance of Classifiers after Oversampling using ADASYN

7. Conclusions and Future Enhancement

The dataset used in this research paper is the recorded performance of individuals in an online game based on phonological awareness, syllabic awareness, auditory discrimination, visual working memory, lexical awareness and many more. Since the dataset had unequal distribution of positive and negative values, the classification accuracy of the different machine learning algorithms would be inaccurate. So three resampling technqiues were used. The first one was undersampling using near-miss algorithm, and oversampling using SMOTE and ADASYN. The two oversampling techniques of SMOTE and ADASYN outperformed the undersampling technique in all aspects of evaluation metrics like accuracy, F1 Score and recall. After applying the undersampling near-miss algorithm, best accuracy was given by SVC classifier with the value of 81.63%. All the other classifiers produced accuracy in the range of 64% to 79.08%. After using the oversampling algorithm SMOTE and using the classifiers produced very good results in all the evaluation metrics of accuracy,CV score, F1 Score and recall. The maximum accuracy was given by RandomForest with a value of 96.37% and closely followed by XGB and GradientBoosting with an accuracy of 95.14%. Decision tree, SVC and ADABoost got an accuracy of 91.26%, 93.36% and 93.48% respectively. Even the values of CV score, F1 and recall were considerably high for all these classifiers. After applying the oversampling technique of ADASYN, RandomForest algorithm generated maximum accuracy of 96.25%. Between the two oversampling techniques, SMOTE algorithm performed slightly better in producing better evaluation metrics than ADASYN. The research work has presented an efficient strategy for predicting the risk of Dyslexia from the given dataset. Application of the sampling techniques like SMOTE and ADASYN increased the classification accuracy. In the future work, hybrid resampling technique can be used to improve the accuracy of classification. Also similar prediction systems can be developed for other learning disabilities like Dyscalculia, Dyspraxia and also for conditions like Autism.

References

[1] Kohli, A., Sharma, S., & Padhy, S. K. (2018). Specific Learning Disabilities: Issues that Remain Unanswered. *Indian journal of psychological medicine*, *40*(5), 399–405. https://doi.org/10.4103/IJPSYM.IJPSYM_86_18

[2] Dyslexia at a glance, International Dyslexia Association, [Last accessed on August 17th, 2021]. Available from https://dyslexiaida.org/dyslexia-at-a-glance/.

[3] O. L. Usman, R. C. Muniyandi, K. Omar and M. Mohamad, "Advance Machine Learning Methods for Dyslexia Biomarker Detection: A Review of Implementation Details and Challenges," in IEEE Access, vol. 9, pp. 36879-36897, 2021, doi: 10.1109/ACCESS.2021.3062709.

[4] Mandatory tests for Dyslexia, Dyslexia Association of India, [Last accessed on August 17, 2021]. Available from https://www.dyslexiaindia.org.in/pdf/test.pdf

[5] H. M. Al-Barhamtoshy and D. M. Motaweh, "Diagnosis of Dyslexia using computation analysis," 2017 International Conference on Informatics, Health & Technology (ICIHT), 2017, pp. 1-7, doi: 10.1109/ICIHT.2017.7899141.

[6] Luz Rello and Miguel Ballesteros. 2015. Detecting readers with dyslexia using machine learning with eye tracking measures. In *Proceedings of the 12th International Web for All Conference* (*W4A '15*). Association for Computing Machinery, New York, NY, USA, Article 16, 1–8. DOI:https://doi.org/10.1145/2745555.2746644

[7] Rello L, Baeza-Yates R, Ali A, Bigham JP, Serra M (2020) Predicting risk of dyslexia with an online gamified test. PLoS ONE 15(12): e0241687.

[8] Dr.A.V. Pradeep Kumar, Dr.G. Vamsi Krishna, Dr.K.Satish Kumar(2018). Diagnosis Children with Dyslexia using Machine Learning technique. International Journal of Pure and Applied Mathematics: Volume 120 No 6, 7305-7320

[9] David, Julie & Balakrishnan, Kannan. (2010). Machine Learning Approach for Prediction of Learning Disabilities in School-Age Children. International Journal of Computer Applications. 9. 10.5120/1432-1931

[10] Nilsson Benfatto M, Öqvist Seimyr G, Ygge J, Pansell T, Rydberg A, Jacobson C (2016) Screening for Dyslexia Using Eye Tracking during Reading. PLoS ONE 11(12): e0165508. https://doi.org/10.1371/journal.pone.0165508

[11] A. Jothi Prabha and R. Bhargavi, ''Predictive model for Dyslexia from fixations and saccadic eye movement events,'' Comput. Methods Programs Biomed., vol. 195, Oct. 2020, Art. no. 105538, doi: 10.1016/j.cmpb.2020.105538.

[12] Liu, Shigang & Zhang, Jun & Xiang, Yang & Zhou, Wanlei & Xiang, Dongxi. (2020). A study of data pre-processing techniques for imbalanced biomedical data classification. International Journal of Bioinformatics Research and Applications. 16. 290. 10.1504/IJBRA.2020.109103

[13] Suresh, K., Thomas, S. V., & Suresh, G. (2011). Design, data analysis and sampling techniques for clinical research. *Annals of Indian Academy of Neurology*, *14*(4), 287–290. https://doi.org/10.4103/0972-2327.91951

[14] Letteri, Ivan & Cecco, Antonio & Dyoub, Abeer & Penna, Giuseppe. (2020). A Novel Resampling Technique for Imbalanced Dataset Optimization

[15] Wenhao Xie, Gongqian Liang, Zhonghui Dong, Baoyu Tan, Baosheng Zhang, "An Improved Oversampling Algorithm Based on the Samples' Selection Strategy for Classifying Imbalanced Data", Mathematical Problems in Engineering, vol. 2019, Article ID 3526539, 13 pages, 2019. https://doi.org/10.1155/2019/3526539

[16] Dataset available at the location : https://doi.org/10.34740/kaggle/dsv/1617514

[17] Mukherjee, M.; Khushi, M. SMOTE-ENC: A Novel SMOTE-Based Method to Generate Synthetic Data for Nominal and Continuous Features. *Appl. Syst. Innov.* 2021, *4*, 18. https://doi.org/10.3390/asi4010018

[18] Taha Muthar Khan, Shengjun Xu, Zullatun Gull Khan, Muhammad Uzair chishti, "Implementing Multilabeling, ADASYN, and ReliefF Techniques for Classification of Breast Cancer Diagnostic through Machine Learning: Efficient Computer-Aided

Diagnostic System", Journal of Healthcare Engineering, vol. 2021, Article ID 5577636, 15 pages, 2021. https://doi.org/10.1155/2021/5577636

[19] Nhlakanipho Michael Mqadi, Nalindren Naicker, Timothy Adeliyi, "Solving Misclassification of the Credit Card Imbalance Problem Using Near Miss", Mathematical Problems in Engineering, vol. 2021, Article ID 7194728, 16 pages, 2021. https://doi.org/10.1155/2021/7194728

[20] Uma R. Salunkhe, Suresh N. Mali, "A Hybrid Approach for Class Imbalance Problem in Customer Churn Prediction: A Novel Extension to Under-sampling", International Journal of Intelligent Systems and Applications(IJISA), Vol.10, No.5, pp.71-81, 2018. DOI: 10.5815/ijisa.2018.05.08

[21] A. Fernández, S. Garcia, F. Herrera, and N. V. Chawla, "SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary," *Journal of Artificial Intelligence Research*, vol. 61, pp. 863–905, 2018.

Authors' Profiles

Ms. Vani Chakraborty has completed her MCA and M.Phil in Computer Applications and currently pursuing PhD in Computational Sciences and IT from Garden City University, Bengaluru. Her research interests include Artificial Intelligence, Data Science, Machine Learning and Deep Learning. She has presented papers in various national and international conferences and journals of repute.

Dr. Meenatchi Sundaram has completed his MCA and M.Phil. He is a Member, Board of Examination for MCA, Bangalore University and many Deemed Universities and Autonomous Colleges. He has compiled a study material on "Microprocessor and its Application" for Bharatidasan University. He is conducting technical skills training for "Campus Connect" Infosys program. He is also a certified ISO 9001:2008 auditor for quality system. He has presented several papers in Mobile Computing. He was Coordinator for M.Phil. Distance education program in the year 2003 and has guided many M.Phil. Students in the field of computer science. His areas of interest are Systems Programming, Assembly Language Programming, Mobile Computing, Data Base Management Systems and Linux Shell Programming.

How to cite this paper: Vani Chakraborty, Meenatchi Sundaram, "An Efficient Smote-based Model for Dyslexia Prediction", International Journal of Information Engineering and Electronic Business(IJIEEB), Vol.13, No.6, pp. 13-21, 2021. DOI: 10.5815/ijieeb.2021.06.02