DIABETES MELLITUS PREDICTION ON CLASS BALANCED DATASET USING XGBOOST ALGORITHM

Harshini Manoharan*, **Dr J Dhilipan****

*Research Scholar, Department of Computer Science,
Faculty of Science and Humanities,
SRM Institute of Science and Technology, Ramapuram Campus,
Bharathi Salai, Ramapuram, Chennai-89, India

**Professor and Head, Department of Computer Science and Applications,
Faculty of Science and Humanities,
SRM Institute of Science and Technology, Ramapuram Campus,
Bharathi Salai, Ramapuram, Chennai-89, India

***Abstract* -** With the advancement in the Information Technology and by the use of various Machine Learning techniques several models were built for predicting DM but majority of the algorithms exhibited an accuracy rate of 70%-90%. This clearly proclaims that still there is a need to build an efficient model capable of classifying distinctly. This paper aims at classifying the subjects into Diabetic and Non-Diabetic classes using the dataset drawn from the National Institute of Diabetic and Digestive and Kidney disease. SMOTE, an oversampling technique which overcomes the class imbalance problem is experimented on the dataset such that the classification dataset does not have a skewed proportion. The class balanced dataset is trained using the XGBoost algorithm, an ensemble technique akin to decision tree that makes use of Gradient Boosting framework out-turn an accuracy score of 97%.

***Keywords* –** Diabetes Mellitus, XGBoost, SMOTE (Synthetic Minority Oversampling Technique)

I. INTRODUCTION

Massive volumes of data in the healthcare sector have high prospects to reduce the medication cost, avert evadable disease, predict outbreaks of epidemic and in general revamp the life quality. Medical practitioners are in an urge to explore more about their patients to pick up warning signs of illness at the earliest. Almost all the patients possess their own Electronic Health Record (EHR) that comes out with high level of security holding details such as medical history, results of laboratory tests and the demographic information. These records drill down the admission of the patients in the hospital by potentially analyzing the data. By leveraging data analysis we can down the error imposed by humans.

Data analysis helps in streamlining the model in making the right decisions. There are several ubiquitous challenges faced in the processing of data analysis regardless of the quantity of data collected for the analysis, meaningfulness of the data, difficulty in handling poor quality data and inaccessibility of data. By performing effective analysis on data potential models can be constructed in predicting the disease. Data analysis are of different type's namely prescriptive, descriptive, diagnostic, and predictive analysis. Descriptive analysis is carried out in clearly figuring the patients affected with diabetes on the dataset collected. Further the model is trained using boosting algorithm to achieve desired result.

In today's world Diabetes Mellitus shortened as DM is a metabolic disease that turns out to be a silent killer and a veteran disease causing more than millions of people thereby leading to various complication such as Diabetic Retinopathy, Diabetic Nephropathy, Diabetic Neuropathy if not treated early. Based on the cause, diabetes comes in different forms such as – Type I Diabetes, Type II Diabetes, Prediabetes and Gestational Diabetes. Type 1 Diabetes (TIDM) labelled as 'Juvenile Diabetes' or in other form insulin dependent diabetes is a kind of disorder where the body disrupts its insulin producing cells present in the pancreas permanently. This type is high-flown mainly in children with the overall rate of 10% in the world's population. In Type 2 Diabetes (T2DM) almost 95% of the population is high-flown mainly in mid aged and older aged groups. In this type, the insulin secreted by pancreas is not brought into play to the fullest and hence the glucose gets accumulated in the blood stream which can be healed by leading a healthy life style with proper medications. Obesity weigh up as the major risk factor of this type. Prediabetes is a condition where the patient has higher blood sugar level than the normal level but not too high to be labelled as Type II Diabetes. Gestational diabetes occurs in women during their gestation. This might cause high risk for the child to develop Prediabetes if not treated properly.

II. RELATED WORKS

Shyamili et al, made a comparative study on the classifying of the diabetic and non-diabetic classes at an early stage by using the same PIMA dataset on various boosting classifiers including Adaboost, GBM, XGBoost and Catboost [1]. Mingqi Li1 et al proposed a methodology for separating the numerical feature and the text feature present in the dataset, RPE and chi square test were used to select the important features and the experimental result showed an accuracy of 80.2%. [2]. Francesco Mercaldoa et al were able to categorize between the diabetic and non-diabetic patients using the PIMA dataset by comparing six different classification algorithm out of which the best classification algorithm was chosen by evaluating the effectiveness of feature vector. The algorithm gave the precision value of 0.757 and the recall value of 0.762 [3]. Liyang Wang et al constructed a questionnaire for their experiment which included personal data of 380 patients like diet, eating habits, exercise, family history which was trained with SVM, KNN, Random forest, and XGBoost out of which XGBoost gave an effective result of 89% accuracy. [4] Roxana Mirshahvalad et al came up with a computer-aided system to leverage the perceptron algorithm performance in predicting the diabetes for the non-diagnostic people performed on three different publicly available dataset thereby reducing the cost incurred in taking the test [5] Aishwarya Mujumdara et al performed K-means clustering on the highly correlated attribute, model was built by implementing SVM,Random forest, Decision tree, Adaboost,LDA, KNN, Naïve Bayes, Perceptron, Gradient boosting and bagging algorithm and further a pipeline model was imposed for improving the accuracy yielding 98.8%.[6].Vandana Rawat et al employed five algorithms on the PIMA dataset for analyzing and predicting the patients and the computed result were found to be 81.77% and 79.69% for bagging and Adaboost technique [7]. Priyadharshini. P presented a paper were that dataset taken from the publicly available UCI repository were trained using the Gradient Boosting algorithm predicting the diabetes with the accuracy of 90% [8]. Amit kumar Dewangan et al constructed an ensemble model by amalgamating the Bayesian classifier along with the Multilayer perceptron to give a new hybrid method producing an accuracy of 81.89%. [9]. D. Vigneswari et al predicted DM by comparing machine learning tree classifiers and achieved an accuracy of 79.31% [10]. J. PradeepKandhasamy et al experimented five machine learning algorithm with the dataset drawn from the UCI repository, and compared the performance of the algorithm with the dataset before preprocessing and dataset after preprocessing [11]. Abdulhakim Salum Hassan et al compared KNN, decision tree and SVM and proved that SVM outperforms other algorithms with the accuracy of 90.23% [12]. Kucharlapati Manoj Varma et al analysed the risk factors of diabetes from the PIMA dataset and compared the result out of which Naïve bayes and SVM gave higher accuracy [13]. V Veena Vijayan et al studied the effect of Principal component analysis and discretization on Naïve Bayes, SVM and Decision tree [14]. Ashok Kumar Dwivedi used six different computational intelligence technique evaluated on eight different performance measure, accuracy of 78% was achieved by logistic regression [15].

Ahmed Saad Hussein et al developed a Novel approach named as ASMOTE, a preprocessing technique which was tested with 44 datasets to fine-tune the newly introduced minority class depending solely on the variation to the original minority class samples.[16] TingtingPan et al presented a paper where sampling methods like Adaptive-SMOTE and Gaussian oversampling methods were compared and tested on 15 dataset proving that Adaptive-SMOTE is more effective than other typical methods for imbalanced dataset [17]. Jia Li et al implemented Random SMOTE integrated with logit to increase the number of minority samples using five UCI imbalanced dataset [18]. Yuanting Yan et al improved the SMOTE based on constructive covering algorithm using a parameter free data cleaning method on 25 imbalanced dataset giving higher metric value [19]. Juanjuan Wang et al experimented locally linear embedding algorithm incorporating in SMOTE algorithm, a novel approach resulted in superior performance than the conventional SMOTE. [20]. BaiyunChen et al RSMOTE has been introduced for imbalanced classification with label noise, that does not rely on any specific noise filter nor any extra parameters [21].

The goal of our research is to help the doctors to make data-driven decision so that appropriate medication can be given to the patients at the earliest and to prevent them from numerous complications. Most of the research proved to give results giving less than 100% by experimenting several traditional methods and hybrid approached. To achieve the desired results at most attention was given to the data analysis phase where the samples from the dataset were class balanced using SMOTE and the model was experimented, trained and tested using the XGBoost Classifier.

III. METHODOLOGY

This section details on the method adopted to discriminate between the Diabetic and Non-Diabetic patients. The entire methodology was completely implemented in python language and the results derived is mentioned in the later section. Fig 1 demonstrates the framework of the proposed methodology.

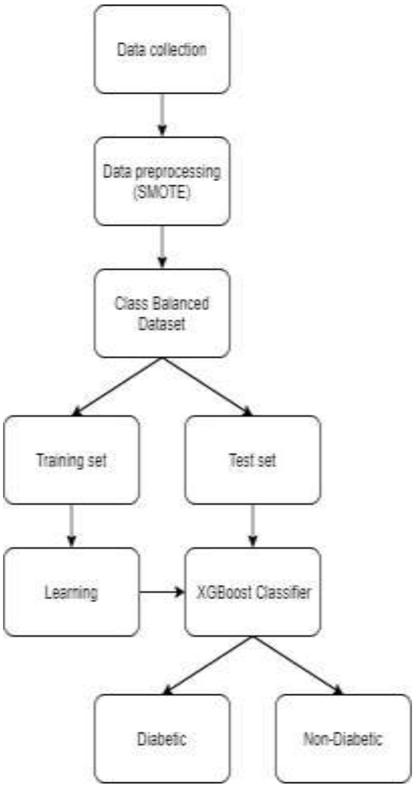

Fig 1. Framework of the proposed methodology

DATASET COLLECTION

The proposed methodology begins with data collection from the National Institute of Diabetic and Digestive and Kidney disease made up of 768 records with eight input attributes for each record and one output attribute as outcome holding the value of either 1 or 0 constituting Diabetic and No-diabetic classes. The description and the data type of the eight input attributes are as follows,

Table 1. Attribute description and the datatype

No	Attribute	Description	Data type
1	Age, expressed in years	Indicates the length of time the person has lived	Numerical value
2	Diabetes pedigree function	Provides history in relatives and genetic relationships	Numerical value
3	Body mass index, quantified in weight in kg/(height in m)ˆ2	Normal – 18.8 to 24.9	Numerical value
4	Diastolic blood pressure, quantified in mm	Pressure in the arteries when the heart rests between the beats Normal : Lower than 80	Numerical value
5	Pregnancy count	Indicates the number of times a women gets pregnancy during her lifespan	Numerical value
6	Plasma glucose an oral glucose tolerance test	Checks how the body moves sugar from blood Normal – 110mg/dl to 160mg/dl	Vector type
7	2-Hour serum insulin, quantified in mu U/ml	Normal less than 150mu u/ml	0/1
8	Triceps skin fold thickness, quantified in mm	Amount of fat present in the body Normal – 2.5mm (men),18.0(women)	Numerical value

DATA PREPROCESSING

Data preprocessing is the supreme part where the data gets transformed or encoded in such a way that the models can easily parse it. The dataset were inspected for any null value or missing values but didn't have any. Heat map were constructed to figure out the magnitude of the attributes in two dimensions. The color variation indicated how the variables are clustered or varies over space. Larger values are represented using dark colors and the lighter color represents the smaller ones.

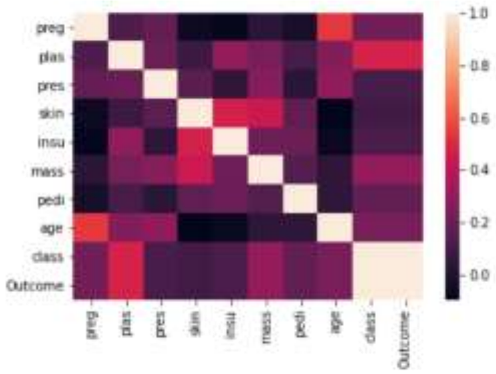

Fig 2. Heat map of the attributes

Synthetic Minority Oversampling Technique (SMOTE) is a kind of resampling method which oversamples the minority class to reduce the biasness of dataset and overcome the class imbalance problem in order to generate the dataset that does not have a skewed proportion. SMOTE computes the variation between the feature vector and the nearest neighbor to generate the synthetic samples. After obtaining the difference any number between 0 and 1 is randomly chosen and multiplied with the variation and added to the feature vector. By using this method we can interpolate several minority classes that lie together. The training dataset comprises of 537 records which is 70% of 768 and the testing data comprises of 231 records which is 30% of 768 records. Among 537 training records number of diabetic patients was 194 and the number of Non-diabetic patients was 343 before oversampling. After applying the sampling technique the number of samples in both the classes were made equal containing 343 diabetic and 343 Non-diabetic patients summing up to 686 records.

Table 2. Number of samples before and after SMOTE

Diabetic/ Non- Diabetic	Before Oversampling	After oversampling
1	194	343
0	343	343
Total	537	686

SMOTE was applied to the remaining 231 testing data which was implemented using python and achieved the following result,

Table 3. Classification report

	Precision	Recall	F1-Score	Support
0	0.97	0.97	0.97	156 / 157
1	0.97	0.97	0.97	73 / 74
Accuracy	-	-	0.97	225 / 231
Macro average	0.97	0.97	0.97	225 / 231
Weighted average	0.97	0.97	0.97	225/ 231

The class balanced dataset is further divided into training set comprising of 70% of dataset making up to 480 records and the test set comprising of 30% of the dataset making up to 206 records for classification through boosting algorithm.

XTREME GRADIENT BOOSTING ALGORITHM

Ensemble learning is a method employed to leverage the performance of the model by coalesced discrete learners thereby enhancing the efficiency and accuracy of the model. Boosting is a type of ensemble learning which utilizes sequential method for producing the weak learners. Xtreme Gradient Boosting, an advanced version of Gradient Boosting is solely developed to focus on the computational speed and model efficiency. XGBoost combines discrete weak leaners into a single strong learner since the rules of all the weak learners are not strong enough to make predictions. Classification is done based on the majority of the voting got from the weak learners. Performance of the model is increased by giving a higher weightage value to the misclassified data. All the misclassified data are considered until the efficiency is boosted as follows,

Step 1 – The base algorithm assigns an equal weightage to each and every sample observation of the dataset.

Step 2 – Higher weightage value to the falsely predicted samples are assigned and passed on to the next base learner for accurate prediction.

Step 3 - The process is repeated until the misclassified data is correctly classified by increasing their weightage in the next iteration.

Weak rules are generated for every single iteration and after collective iteration the weak learners are coalesced together which predicts the outcome accurately.

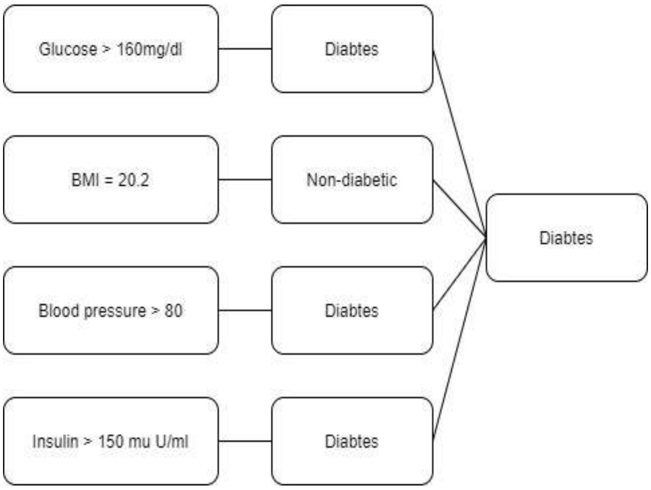

Fig 3. Weak leaners for highly correlated attributes

The first step in fitting XGBoost to the training data is to make the initial prediction. Probability value of 0.5 which is the default value is taken indicating that the patient has 50% change of being diabetic. Values greater than 0.5 indicates diabetic patients and those that are below 0.5 is labelled as Non-diabetic patients. Residuals for each record is computed using the following formula,

Residual = Observed value (O) – Predicted value (E) (1)

The minimum number of residuals in each leaf is determined by calculating the cover value.

Cover = ∑[Previous probability * (1- Previous probability)] (2)

If the cover value is 1, which is the default value we will have only root node since XGBoost requires tree larger than the root. So in order to prevent we set the minimum value of Cover to 0.

Similarity Score = (∑Residuals $_i$)^2 / ∑[Previous probability * (1- Previous probability)]+ λ (3)

Where λ is the regularization parameter that minimizes the sensitivity to related observations. The value of λ is set to 0 that reduces the similarity score which ultimately makes leaves easier to prune. Gain is calculated once the similarity score is calculated for all the nodes after splitting the tree using the formula,

Gain = Left $_{\text{Similarity}}$ + Right $_{\text{Similarity}}$ – Root $_{\text{Similarity}}$ (4)

The tree is pruned by calculating the variation between the Gain associated with the lowest branch and γ. If the difference is a negative number then the tree is pruned.

Gain – γ = If positive, then do not prune (5)

If negative, then prune

Finally the output value for each node of the tree is calculated using the formula,

Output = Residuals / p(1-p) + λ (6)

Just like other boosting methods, XGBoost makes new predictions by starting off with the initial probability however this must be converted to a log(odds) value

Odds = P/1-P (7)

By applying log on both sides we get,

Log $_{\text{odds}}$ = log (P/1-P) (8)

So when p=0.5 the log$_{\text{odds}}$ = 0

Log$_{\text{odds}}$ of the initial prediction to the output of the tree, scaled by the learning rate ε = 0.3, which is the default value. The new value for the observation is calculated using,

Log$_{\text{(odds)}}$ Prediction = Log $_{\text{odds}}$ + (ε * Output) (9)

Converting the log(odds) value into a probability we plug it into the Logistic function

Probability = e $^{\text{log(odds)}}$ / 1 + e $^{\text{log(odds)}}$ (10)

The process is repeated for the next tree which has a smaller residual value.

IV. CLASSIFIER PERFORMANCE MEASURE EVALUATIONS

XGBoost algorithm was applied on the class balanced dataset containing 686 records using Python language. The accuracy of the model was estimated and found to be 97%. Confusion matrix or the error matrix is used to evaluate the performance of the classifier by counting the number of correct and incorrect values broken down by each class. Given below is the resultant matrix obtained from the balanced dataset. From the confusion matrix Accuracy, Recall, Precision, F-measure, Sensitivity, Specificity was computed and resultant value of 0.97 were derived indicating 97% efficiency of the model.

Table 4. Confusion Matrix

	Positive (1)	Negative (0)
Positive (1)	670	16
Negative (0)	16	670

Accuracy of a classifier represents the quantity of correctly diagnosed patients with diabetes and non-diabetes. Precision is the measure of the number of positives correctly classified from all the positives. Recall measures the proportion of actual diabetic patients that were correctly computed. F1-score measures the performance of the models classification ability by calculating the harmonic mean of precision and recall. Sensitivity is the ratio between the correctly identified diabetic patients to the actual diabetic cases. Specificity is the ratio between the correctly identified non-diabetic patients to the actual non-diabetic cases. The values are tabulated below,

Table 5. XGBoost Classification Report

	Precision	Recall	F1-Score
0	**0.97**	**0.97**	**0.97**
1	**0.97**	**0.97**	**0.97**
Accuracy	**-**	**-**	**0.97**
Macro average	**0.97**	**0.97**	**0.97**
Weighted average	**0.97**	**0.97**	**0.97**

V. CONCLUSION AND FUTURE ENHANCEMENT

Several Machine learning algorithm were employed to differentiate between the Diabetic and Non-Diabetic patients which includes SVM, KNN, Random Forest, Linear Regression, Naïve Bayes. Most of the papers experimented to combine the algorithms so that the accuracy of the model is leveraged. A key contribution of this paper is to clearly discriminate between the Diabetic and Non-diabetic patients with the XGBoost algorithm performed on the class balanced dataset. The model showed a promising result in predicting the disease. Our interpretation shows that the machine learning model based on the boosting algorithm were able to procure an accuracy of 97%. Patients with Diabetes Mellitus are identified by the models based largely on their attributes (age, blood pressure, BMI, pregnancy, pedigree function, insulin level, glucose, skin thickness). As a part of our future works we are planning to explore the complications of gestational diabetes towards the development of accurate model.

REFERENCES

[1] Shyamili V and Shilpa Ankalak, "Boosting Classifiers in Diabetes `Disease Prediction", International Journal of Emerging Trends in Engineering Research, Volume 8. No. 7, July 2020, ISSN 2347 – 3983.

[2] Mingqi Li1, Xiaoyang Fu1, and Dongdong Li, "Diabetes Prediction Based on XGBoost Algorithm", IOP Conf. Series: Materials Science and Engineering 768 (2020) 072093, doi:10.1088/1757-899X/768/7/07209.

[3] Francesco Mercaldoa, Vittoria Nardoneb, Antonella. "Diabetes Mellitus Affected Patients Classification and Diagnosis through Machine Learning Techniques", International Conference on Knowledge Based and Intelligent Information and EngineeringSystems, KES2017, 6-8 September 2017, Marseille, France, Procedia Computer Science 112 (2017) 2519–2528.

[4] Liyang Wang, Xiaoya Wang, Angxuan Chen, Xian Jin and Huilian Che, "Prediction of Type 2 Diabetes Risk and Its Effect Evaluation Based on the XGBoost Model", Healthcare 2020, 8, 247; doi:10.3390/healthcare8030247.

[5] Roxana Mirshahvalad and Nastaran Asadi Zanjani, "Diabetes prediction using ensemble perceptron algorithm", 2017 9th International Conference on Computational Intelligence and Communication Networks, 978-1-5090-5001-7/17/$31.00 ©2017 IEEE, DOI 0.1109/CICN.2017.42

[6] Aishwarya Mujumdara, Dr. Vaidehi V, "Diabetes Prediction using Machine Learning Algorithms", International Conference On Recent Trends In Advanced Computing 2019, ICRTAC 2019, Procedia Computer Science 165 (2019) 292–299.

[7] Vandana Rawat and Suryakant, "A Classification System for Diabetic Patients with Machine Learning Techniques", International Journal of Mathematical, Engineering and Management Sciences Vol. 4, No. 3, 729–744, 2019.

[8] Priyadharshini. P, "Prediction Of Diabetes Mellitus Using Xgboostgradient Boosting", International Journal Of Advances In Science Engineering And Technology, ISSN(P): 2321 –8991, ISSN(E): 2321 –9009 Vol-5, Iss-4, Spl. Issue-2 Dec-2017.

[9] Amit kumar Dewangan, Pragati Agrawal, "Classification of Diabetes Mellitus Using Machine Learning Techniques", International Journal of Engineering and Applied Sciences (IJEAS) ISSN: 2394-3661, Volume-2, Issue-5, May 2015.

[10] Ahmed Saad Hussein, Tianrui Li, Chubato Wondaferaw Yohannese, Kamal Bashir, "A-SMOTE: A New Preprocessing Approach for Highly Imbalanced Datasets by Improving SMOTE", International Journal of Computational Intelligence Systems, Volume 12, Issue 2, 2019, Pages 1412 – 1422.

[11] D. Vigneswari, N. Komal Kumar, V. Ganesh Raj, A. Gugan, S. R. Vikash, "Machine Learning Tree Classifiers in Predicting Diabetes Mellitus", Publisher: IEEE, 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS).

[12] J. PradeepKandhasamy and S.Balamurali, "Performance Analysis of Classifier Models to Predict Diabetes Mellitus",Procedia Computer Science 47(2015)45–51, Volume 47, 2015, https://doi.org/10.1016/j.procs.2015.03.182.

[13] Abdulhakim Salum Hassan, I. Malaserene, A. Anny Leema, "Diabetes Mellitus Prediction using Classification Techniques", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-9 Issue-5, March 2020.

[14] Kucharlapati Manoj Varma, Dr B S Panda, "Comparative analysis of Predicting Diabetes Using Machine Learning Techniques", Journal of Emerging Technologies and Innovative Research (JETIR), June 2019, Volume 6, Issue 6.

[15] V Veena Vijayan and C Anjali, "Decision support systems for predicting diabetes mellitus — A Review", Publisher: IEEE, Proceedings of 2015 Global Conference on Communication Technologies(GCCT 2015).

[16] Ashok Kumar Dwivedi, Analysis of computational intelligence techniques for diabetes mellitus prediction", The Natural Computing Applications Forum 2017, DOI 10.1007/s00521-017-2969-9

[17] TingtingPan, JunhongZhao, WeiWuJieYang, "Learning imbalanced datasets based on SMOTE and Gaussian distribution", Information Sciences, Volume 512, February 2020, Pages 1214-1233.

[18] Jia Li, Hui Li, Jun-Ling Yu, "Application of Random-SMOTE on Imbalanced Data Mining", 2011 Fourth International Conference on Business Intelligence and Financial Engineering, 03 January 2012, DOI: 10.1109/BIFE.2011.25.

[19] Yuanting Yan, Ruiqing Liu, Zihan Ding, Xiuquan Du, Jie Chen, Yanping Zhang, "A Parameter-Free Cleaning Method for SMOTE in Imbalanced Classification", IEEE Access (Volume: 7), DOI: 10.1109/ACCESS.2019.2899467.

[20] Juanjuan Wang, Mantao Xu, Hui Wang, Jiwu Zhang, "Classification of Imbalanced Data by Using the SMOTE Algorithm and Locally Linear Embedding", Publisher: IEEE, DOI: 10.1109/ICOSP.2006.345752.

[21] BaiyunChen, ShuyinXia, ZizhongChen, BingguiWang, GuoyinWang, "RSMOTE: A self-adaptive robust SMOTE for imbalanced problems with label noise", Information Sciences, PII: S0020-0255(20)31004-5 DOI: https://doi.org/10.1016/j.ins.2020.10.013 Reference: INS 15927.

.