Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 12, Issue 6, July, 2021: 7697-7715

Research Article

Overhead Cross Section Sampling Machine Learning based Cervical Cancer Risk Factors Prediction

A. Peter Soosai Anandaraj[1*], M. Shyamala Devi[2], J. Amutharaj[3], M. Dineshkumar[4]

Abstract

Most forms of human papillomavirus can create alterations on a woman's cervix that can lead to cervical cancer in the long run, while others can produce genital or epidermal tumors. Cervical cancer is a leading cause of morbidity and mortality among women in low- and middle-income countries. The prediction of cervical cancer still remains an open challenge as there are several risk factors affecting the cervix of the women. By considering the above, the cervical cancer risk factor dataset from KAGGLE data warehouse is executed for predicting the cervical cancer risk classes. The cervical cancer data set is normalised with incomplete data and Pattern Calibration. Secondly, the interpretive data analysis is carried out, and the target feature's dispersion of the cervical cancer risk is visualised. Thirdly, several classifiers are fitted to the unprocessed data set, and the performance is measured with pre and post feature scaling. Fourth, oversampling methodologies are applied to the pre - processed data set. Fifth, the oversampled dataset by differment methods are applied to all the classifiers and the performance is compared with pre and post feature scaling. Sixth, Precision, recall, F-score, accuracy, and running time are some of the metrics used in performance analysis. The code is written in Python and executed with Anaconda Navigator on the Spyder framework. The findings of the experiments reveal that the Random forest classifier tends to sustain 96% accuracy pre and post scaling for unporocessed dataset. Similarly the same classifier tends to sustain 98% accuracy for all the oversampling techniques.

Keywords: *Machine learning, scaling, oversampling, precision, accuracy, classification*

[1,2] Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Department of Computer Science & Engineering

[3,4] RajaRajeswari College of Engineering, Department of Information Science and Engineering

Email id: anandsiriya@gmail.com

Introduction

In the world, cervical cancer is the main cause of mortality among women. The lack of awareness of socio-contextual variables of access to screening and prevention opportunities contributes to the high prevalence of cervical cancer. Comprehensive cervical cancer prevention programmes, which include protective factors through human papillomavirus (HPV) immunisation and preventive services through successful therapy of cervical cancer precursors, are largely preventable. Women even now face additional barriers to participating in current cervical cancer screening programmes. Weak health systems, insufficient funding and staff to establish routine screening programmes, high screening expenses, a lack of awareness and education about current programmes, and late presentation and diagnosis are only a few of them.

Related works

Background

The data given in this work highlight the importance of individual, familial, communal, and structural factors that might help or impede women at risk in the study group's access to cervical cancer screening. These factors should be taken into account when developing plans to improve women's access to cervical cancer screening in Cameroon and other low-income countries [1]. Due to its imbalance and large number of missing values, the dataset requires extensive data pre-processing. The performance of classification strategies was assessed using 10-fold cross validation, using accuracy, precision, and recall as assessment metrics. The efficiency strength of all classification methods was determined using a correlation coefficient [2]. This research examines cervical cancer prediction utilising data mining approaches such as decision tree, decision forest, and decision jungle algorithms, as well as performance assessment using the AOC curve, accuracy, specificity, and sensitivity. The results were authenticated using a 10-fold cross-validation procedure, and the decision tree provided excellent outcome [3]. This study present a strategy for systematically analysing costs and effectiveness in order to translate clinical trial findings into screening programme implementation guidelines that optimise benefits in the real world. Evaluating total screening costs can be misleading because the resources spent on certain programme activities could have a direct influence towards numerous parameters and features [4]. However, depending on the degree of economic development and associated social and lifestyle factors, the most often diagnosed cancer and the primary cause of cancer death differ significantly across

countries and within each country. It is important that elevated cancer registry information, which is used to design and implement scientific proof cancer control initiatives, is really not available [5]. Data mining has the potential to improve the efficiency and effectiveness of healthcare systems by allowing them to utilise data more effectively. As a result, It improves health and lowers costs. This article examines a variety of data sources. Techniques like classification and clustering are used in data mining. In the health domain, there is a connection and a regression. It also emphasizes Information Retrieval applications, problems, and current challenges in healthcare [6]. This study conveys that individuals are likely to experience recurrent occurrences with contemporaneous variables, two data mining strategies were investigated. After adjusting for the four most relevant prognostic markers, including pStage, Pathologic T, cell type, and RT target Summary. Finally, patients should be categorised by these prognostic markers in treatment trials, and precise measurement of status could improve result [7]. Given risk patterns from individual medical data, we describe a computationally automated technique for predicting the result of a patient biopsy. We present a machine learning strategy for optimising dimensionality reduction and classification models simultaneously and completely supervised.

To make the lesson easier, create a model that highlights relevant qualities in a low-dimensional space [8]. This paper explain a software suite we created to construct and make freely accessible several of these prediction methods, as well as a computational strategy based on Latent Semantic Indexing (LSI), which searches for semantically comparable genes using both inferred and available annotations. BioAnnotationPredictor is a computer programme that forecasts specific gene functionalities [9]. Those who might have a hysterectomy no longer need to be screened unless they have high-grade dysplasia. Although the value of clinical exam in women with symptoms related to the upper reproductive tract is undisputed, several professional organizations have endorsed less than yearly cervical screening in healthy women, raising the question of whether yearly abdominal examination is beneficial to asymptomatic individuals [10]. İn this work, Powerful Data Mining methods are used in this study to predict if a cervix is normal or cancerous. Data mining is critical for prediction, particularly in the medical field. This approach is used to introduce the Classification and Regression Tree method, the Random Forest Tree algorithm, and RFT with K-means learning for predicting if a cervix is benign or cancerous [11].

Proposed Work

For the implementation, the Cervical cancer risk dataset with 35 independent factors and 1 dependent variable was employed [12]. The following contributions are used to predict cervical cancer. The systematic flow of the work is shown in figure.1.

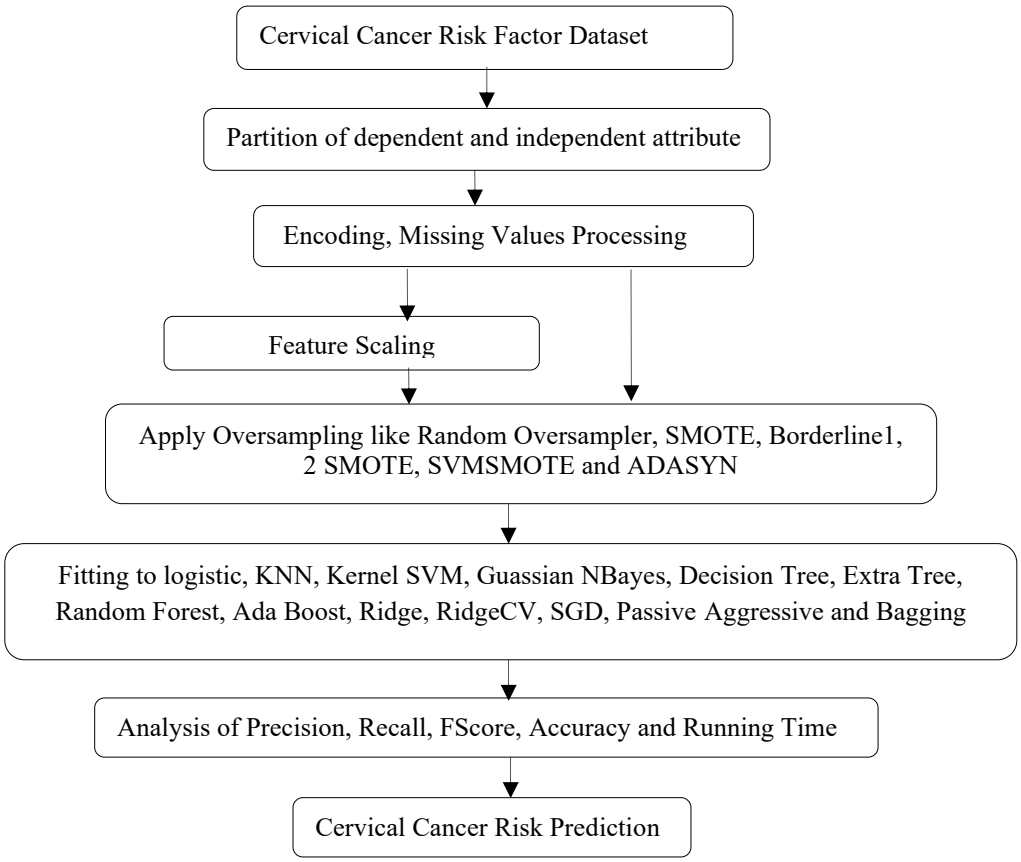

Figure 1. Overall Workflow of the system

(i) The cervical cancer data set is normalised with incomplete data and Pattern Calibration.

(ii) Secondly, the interpretive data analysis is carried out, and the target feature's dispersion of the cervical cancer risk is visualised.

(iii) Thirdly, several classifiers are fitted to the unprocessed data set, and the performance is measured with pre and post feature scaling.

(iv) The target “Biopsy” feature is found to be imbalanced by having 93.6% of healthy people and 6.4% of cervical cancer people. This imbalanced target feature is proposed to be balanced with oversampling methods like random oversampling, SMOTE, SVMSMOTE, Borderline 1,2 and ADASYN techniques.

(v) Fourth, oversampling methodologies are applied to the pre - processed data set.

(vi) Fifth, the oversampled dataset by differment methods are applied to all the classifiers and the performance is compared with pre and post feature scaling.

(vii) Sixth, Precision, recall, F-score, accuracy, and running time are some of the metrics used in performance analysis.

Exploratory Data Analysis

The Cervical cancer risk dataset with 35 independent factors and 1 dependent variable is used for implementation [13-14]. The dataset conatins of 858 patients clinical details with 36 features (Age, Number of sexual partners, First sexual intercourse, Num of pregnancies, Smokes, Smokes (years), Smokes (packs/year), Hormonal Contraceptives, Hormonal Contraceptives (years), IUD, IUD (years), STDs, STDs (number), STDs:condylomatosis, STDs: cervical condylomatosis, STDs:vaginal condylomatosis, STDs:vulvo-perineal, condylomatosis, STDs:syphilis,STDs:pelvic inflammatory disease, STDs:genital herpes, STDs:molluscum contagiosum, STDs:AIDS, STDs:HIV, STDs:Hepatitis B, STDs:HPV, STDs: Number of diagnosis, STDs: Time since first diagnosis, STDs: Time since last diagnosis, Dx:Cancer, Dx:CIN, Dx:HPV, Dx, Hinselmann, Schiller, Citology) and 1 Target “Biopsy”[15-16]. The code is written in Python and evaluated using Anaconda Navigator with the Spyder Interface. The data set is separated into two parts: training with 80% of raw dataset and testing with 20% of raw datase[17-18]t. Figure. 2. shows the target feature analysis and Figure. 3 shows the target distribution and found to be non-sampled with 93.6% of healthy people and 6.4% of cervical cancer people [19-21].

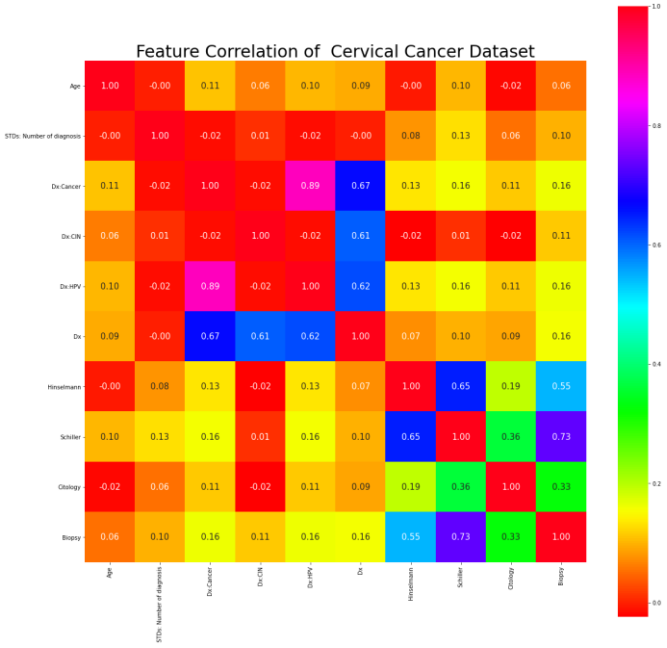

Figure 2. Correlation Analysis of Cervical Cancer Dataset

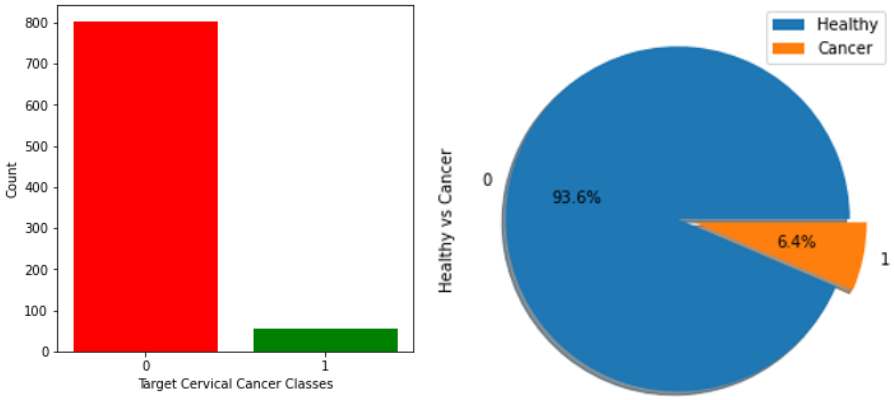

Figure 3. Target Feature Analysis of Dataset

Implementation and Discussion

The unprocessed dataset is applied to all the classifiers and the efficiency metrics is compared pre and post feature scaling and is shown in Table 1 and Table 2, the accuracy and running time comparison is shown in Figure. 4.

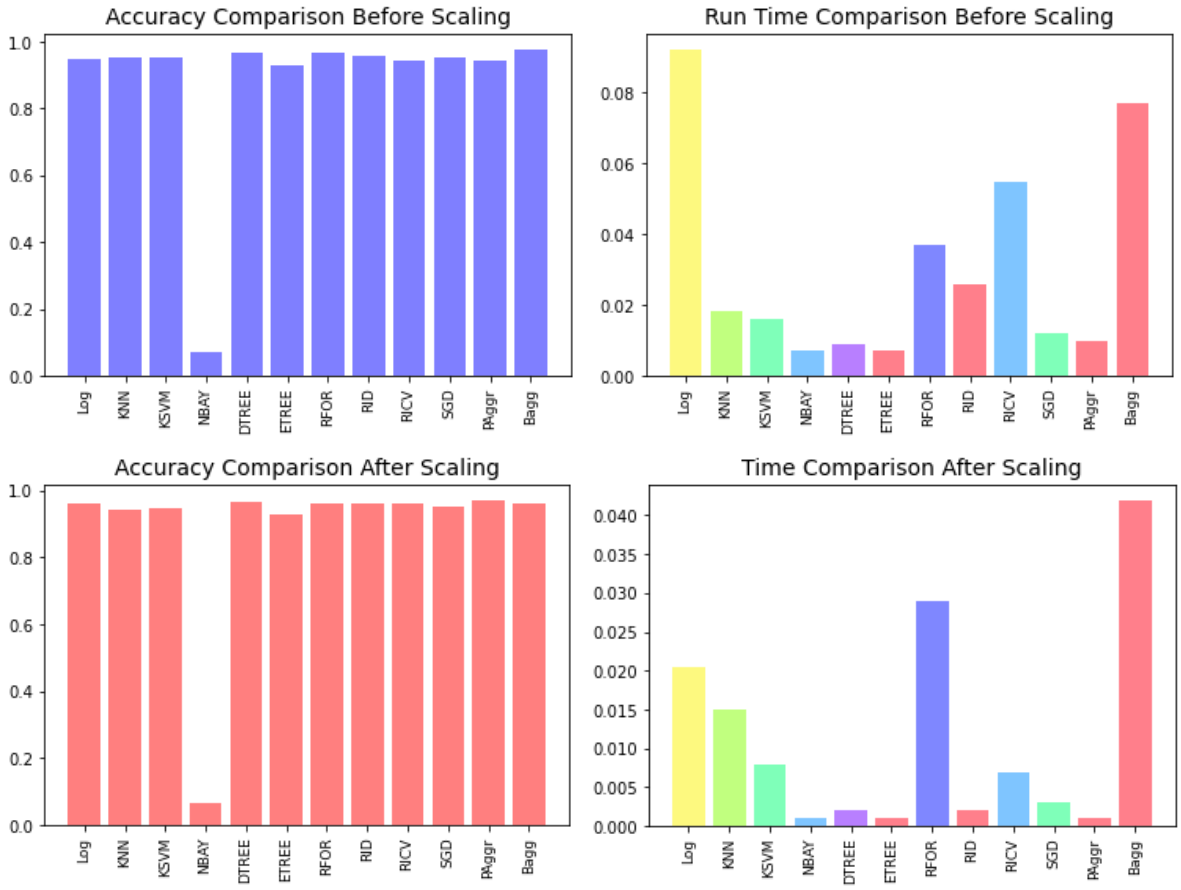

Figure 4. Run Time and Accuacry analysis pre and post scaling of unprocessed dataset

Table 1

Efficiency metrics of unprocessed dataset pre scaling

Classifier	Precision	Recall	FScore	Accuracy	Running Time (ms)
Logistic	0.94	0.95	0.94	0.95	0.09
KNN	0.91	0.95	0.93	0.95	0.02
KSVM	0.91	0.95	0.93	0.95	0.02
GNBayes	0.96	0.07	0.05	0.07	0.01
DTree	0.97	0.97	0.97	0.97	0.01
ETree	0.94	0.93	0.93	0.93	0.01
RForest	0.96	0.97	0.96	0.97	0.04
AdaBoost	0.96	0.96	0.96	0.96	0.03
Ridge	0.93	0.94	0.93	0.94	0.05
RidgeCV	0.91	0.95	0.93	0.95	0.01
SGD	0.93	0.94	0.94	0.94	0.01
PAggress	0.98	0.98	0.98	0.98	0.08
Bagging	0.94	0.95	0.94	0.95	0.09

Table 2

Efficiency metrics of unprocessed dataset post scaling

Classifier	Precision	Recall	FScore	Accuracy	Running Time (ms)
Logistic	0.96	0.96	0.96	0.96	0.02
KNN	0.91	0.94	0.92	0.94	0.01
KSVM	0.93	0.95	0.94	0.95	0.01
GNBayes	0.96	0.06	0.04	0.06	0.00
DTree	0.97	0.97	0.97	0.97	0.00
ETree	0.94	0.93	0.93	0.93	0.00
RForest	0.95	0.96	0.95	0.96	0.03
AdaBoost	0.96	0.96	0.96	0.96	0.00
Ridge	0.96	0.96	0.96	0.96	0.01
RidgeCV	0.94	0.95	0.94	0.95	0.00
SGD	0.97	0.97	0.97	0.97	0.00
PAggress	0.95	0.96	0.95	0.96	0.04
Bagging	0.96	0.96	0.96	0.96	0.02

Quantative Analysis with oversampling

The target “Biopsy” feature is found to be imbalanced by having 93.6% of healthy people and 6.4% of cervical cancer people. This imbalanced target feature is proposed to be balanced with oversampling methods like random oversampling, SMOTE, SVMSMOTE, Borderline 1,2 and ADASYN techniques. The oversampled dataset distribution after all the oversampling methods are shown in Figure.5. The unprocessed dataset is processed with random o versampling and then passed to all classifiers to examine the metrics and is shown in Figure. 6. The comparison of the efficiency metrics re and post scaling is shown in Table. 3 and Table. 4.

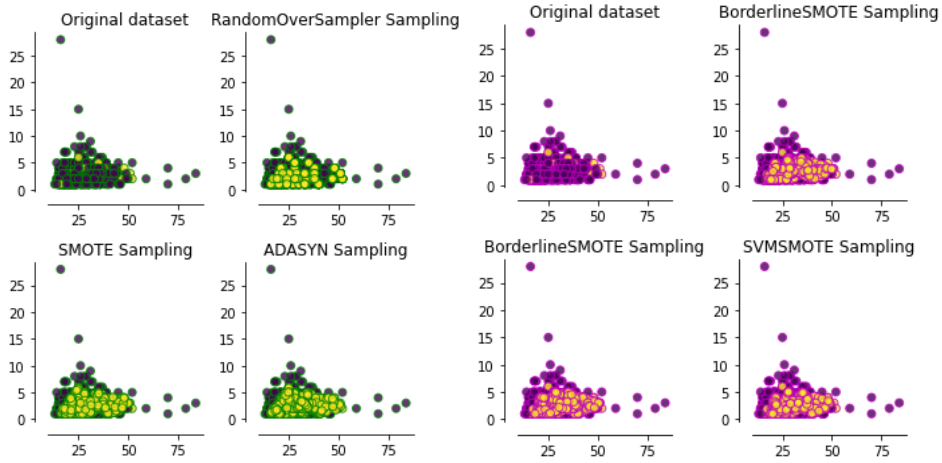

Figure 5. Data distribution after undersampling methods

Table 3

Random oversampling Metrics pre feature scaling

Classifier	Precision	Recall	FScore	Accuracy	Running Time (ms)
Logistic	0.94	0.94	0.94	0.94	0.06
KNN	0.91	0.89	0.89	0.89	0.03
KSVM	0.84	0.84	0.84	0.84	0.11
GNBayes	0.75	0.50	0.34	0.50	0.01
DTree	0.97	0.97	0.97	0.97	0.01
ETree	0.97	0.97	0.97	0.97	0.01
RForest	0.98	0.98	0.98	0.98	0.05
AdaBoost	0.97	0.97	0.97	0.97	0.68
Ridge	0.95	0.95	0.95	0.95	0.01
RidgeCV	0.95	0.95	0.95	0.95	0.01
SGD	0.95	0.95	0.95	0.95	0.02
PAggress	0.93	0.92	0.92	0.92	0.02
Bagging	0.98	0.98	0.98	0.98	0.17

Table 4

Random oversampling Metrics post feature scaling

Classifier	Precision	Recall	FScore	Accuracy	Running Time (ms)
Logistic	0.94	0.94	0.94	0.94	0.03
KNN	0.96	0.96	0.96	0.96	0.04
KSVM	0.95	0.95	0.95	0.95	0.02
GNBayes	0.75	0.50	0.34	0.50	0.00
DTree	0.97	0.97	0.97	0.97	0.00
ETree	0.97	0.97	0.97	0.97	0.00
RForest	0.98	0.98	0.98	0.98	0.04
AdaBoost	0.97	0.97	0.97	0.97	0.11
Ridge	0.95	0.95	0.95	0.95	0.00
RidgeCV	0.95	0.95	0.95	0.95	0.01
SGD	0.94	0.93	0.93	0.93	0.01
PAggress	0.94	0.94	0.94	0.94	0.00
Bagging	0.98	0.98	0.98	0.98	0.04

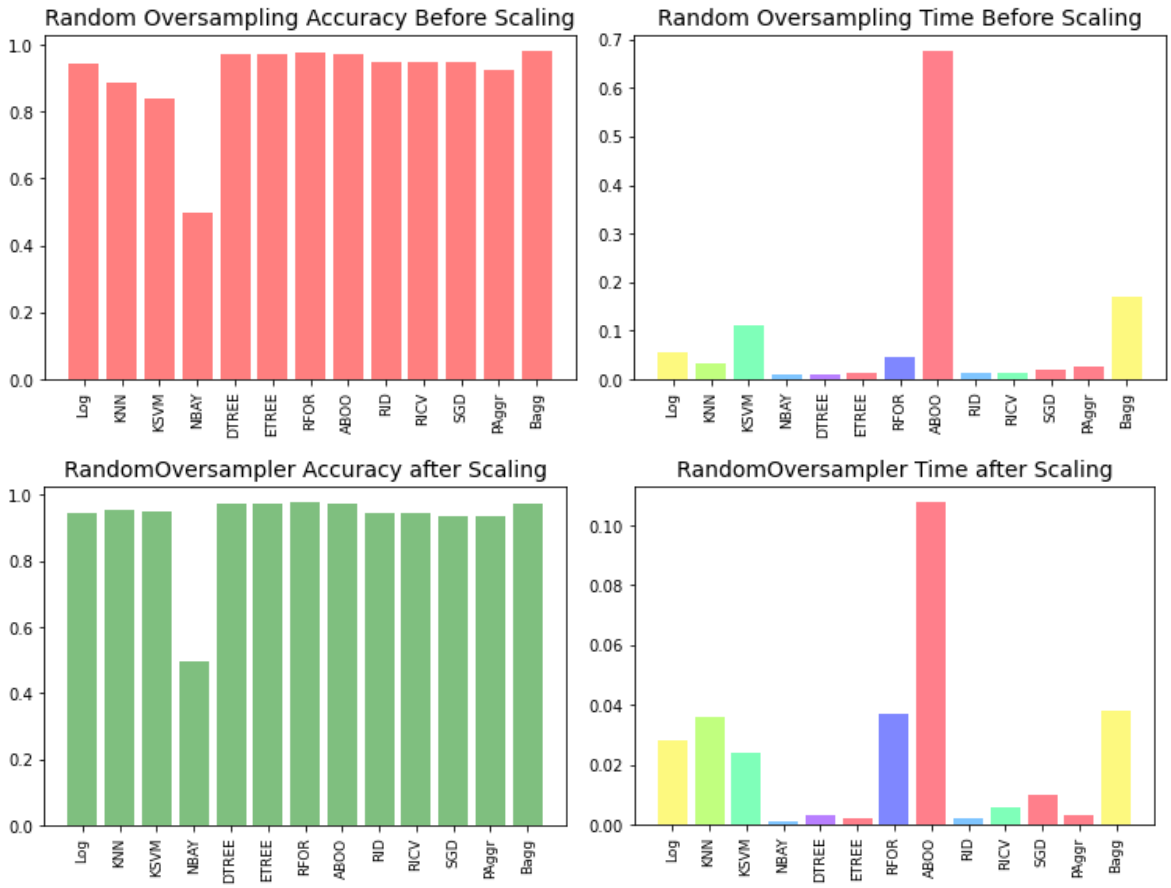

Figure 6. Random Oversampling Accuracy and Time Analysis pre and post feature scaling

The unprocessed dataset is processed with SMOTE oversampling and then passed to all classifiers to examine the metrics and is shown in Figure. 7. The comparison of the efficiency metrics re and post scaling is shown in Table. 5 and Table. 6.

Table 5

SMOTE oversampling metrics pre feature scaling

Classifier	Precision	Recall	FScore	Accuracy	Running Time (ms)
Logistic	0.84	0.83	0.83	0.83	0.06
KNN	0.85	0.83	0.82	0.83	0.03
KSVM	0.76	0.72	0.71	0.72	0.11
GNBayes	0.75	0.50	0.34	0.50	0.01
DTree	0.95	0.95	0.95	0.95	0.01
ETree	0.93	0.93	0.93	0.93	0.01
RForest	0.97	0.97	0.97	0.97	0.05
AdaBoost	0.96	0.96	0.96	0.96	0.23
Ridge	0.85	0.83	0.83	0.83	0.01
RidgeCV	0.85	0.83	0.83	0.83	0.01
SGD	0.84	0.82	0.82	0.82	0.02
PAggress	0.85	0.83	0.83	0.83	0.02
Bagging	0.96	0.96	0.96	0.96	0.09

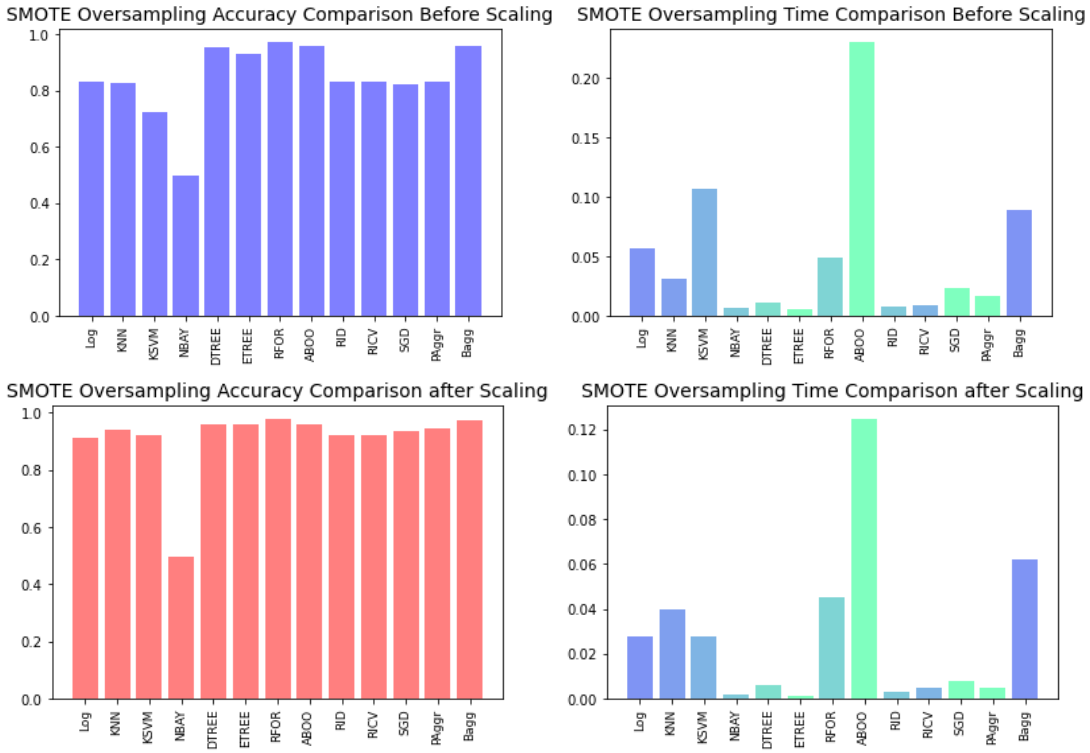

Figure 7. SMOTE Oversampling Accuracy and Time Analysis pre and post feature scaling

Table 6

SMOTE oversampling metrics post feature scaling

Classifier	Precision	Recall	FScore	Accuracy	Running Time (ms)
Logistic	0.91	0.91	0.91	0.91	0.03
KNN	0.94	0.94	0.94	0.94	0.04
KSVM	0.92	0.92	0.92	0.92	0.03
GNBayes	0.75	0.50	0.34	0.50	0.00
DTree	0.96	0.96	0.96	0.96	0.01
ETree	0.96	0.96	0.96	0.96	0.00
RForest	0.98	0.98	0.98	0.98	0.04
AdaBoost	0.96	0.96	0.96	0.96	0.12
Ridge	0.92	0.92	0.92	0.92	0.00
RidgeCV	0.92	0.92	0.92	0.92	0.00
SGD	0.94	0.93	0.93	0.93	0.01
PAggress	0.95	0.94	0.94	0.94	0.00
Bagging	0.97	0.97	0.97	0.97	0.06

The unprocessed dataset is processed with SVM SMOTE oversampling and then passed to all classifiers to examine the metrics and is shown in Figure. 8. The comparison of the efficiency metrics re and post scaling is shown in Table. 7 and Table. 8.

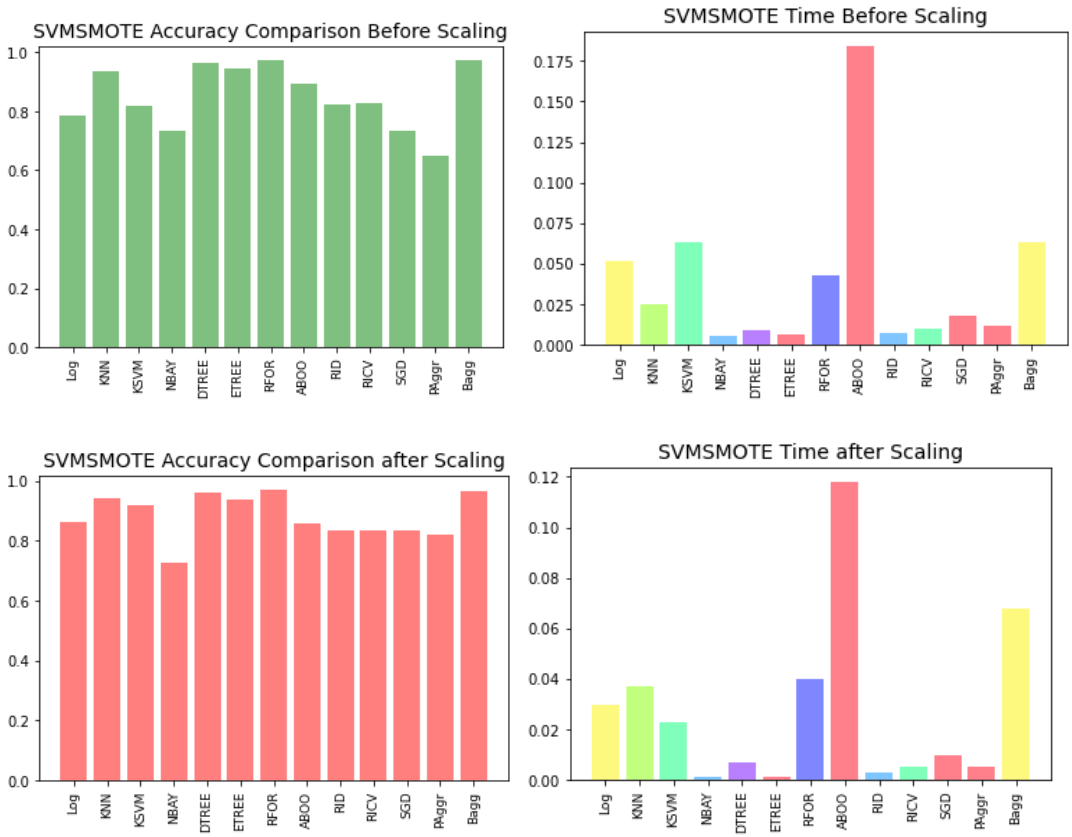

Figure 8. SVM SMOTE Oversampling Accuracy and Time Analysis pre and post scaling

Table 7

SVM SMOTE oversampling metrics pre feature scaling

Classifier	Precision	Recall	FScore	Accuracy	Running Time (ms)
Logistic	0.88	0.87	0.87	0.87	0.05
KNN	0.90	0.90	0.90	0.90	0.02
KSVM	0.75	0.67	0.61	0.67	0.06
GNBayes	0.76	0.43	0.28	0.43	0.00
DTree	0.96	0.96	0.96	0.96	0.01
ETree	0.95	0.95	0.95	0.95	0.01
RForest	0.96	0.96	0.96	0.96	0.04
AdaBoost	0.94	0.93	0.93	0.93	0.18
Ridge	0.88	0.88	0.88	0.88	0.01
RidgeCV	0.89	0.89	0.88	0.89	0.01
SGD	0.88	0.89	0.88	0.89	0.02
PAggress	0.88	0.88	0.88	0.88	0.01
Bagging	0.96	0.96	0.96	0.96	0.06

Table 8

SVM SMOTE oversampling metrics post feature scaling

Classifier	Precision	Recall	FScore	Accuracy	Running Time (ms)
Logistic	0.96	0.96	0.96	0.96	0.03
KNN	0.95	0.95	0.95	0.95	0.04
KSVM	0.97	0.97	0.97	0.97	0.02
GNBayes	0.75	0.50	0.34	0.50	0.00
DTree	0.98	0.98	0.98	0.98	0.01
ETree	0.98	0.98	0.98	0.98	0.00
RForest	0.98	0.98	0.98	0.98	0.04
AdaBoost	0.98	0.98	0.98	0.98	0.12
Ridge	0.95	0.95	0.95	0.95	0.00
RidgeCV	0.95	0.95	0.95	0.95	0.00
SGD	0.95	0.95	0.95	0.95	0.01
PAggress	0.93	0.93	0.93	0.93	0.01
Bagging	0.98	0.98	0.98	0.98	0.07

The unprocessed dataset is processed with Borderline 1 oversampling and then passed to all classifiers to examine the metrics and is shown in Figure. 9. The comparison of the efficiency metrics re and post scaling is shown in Table. 9 and Table. 10.

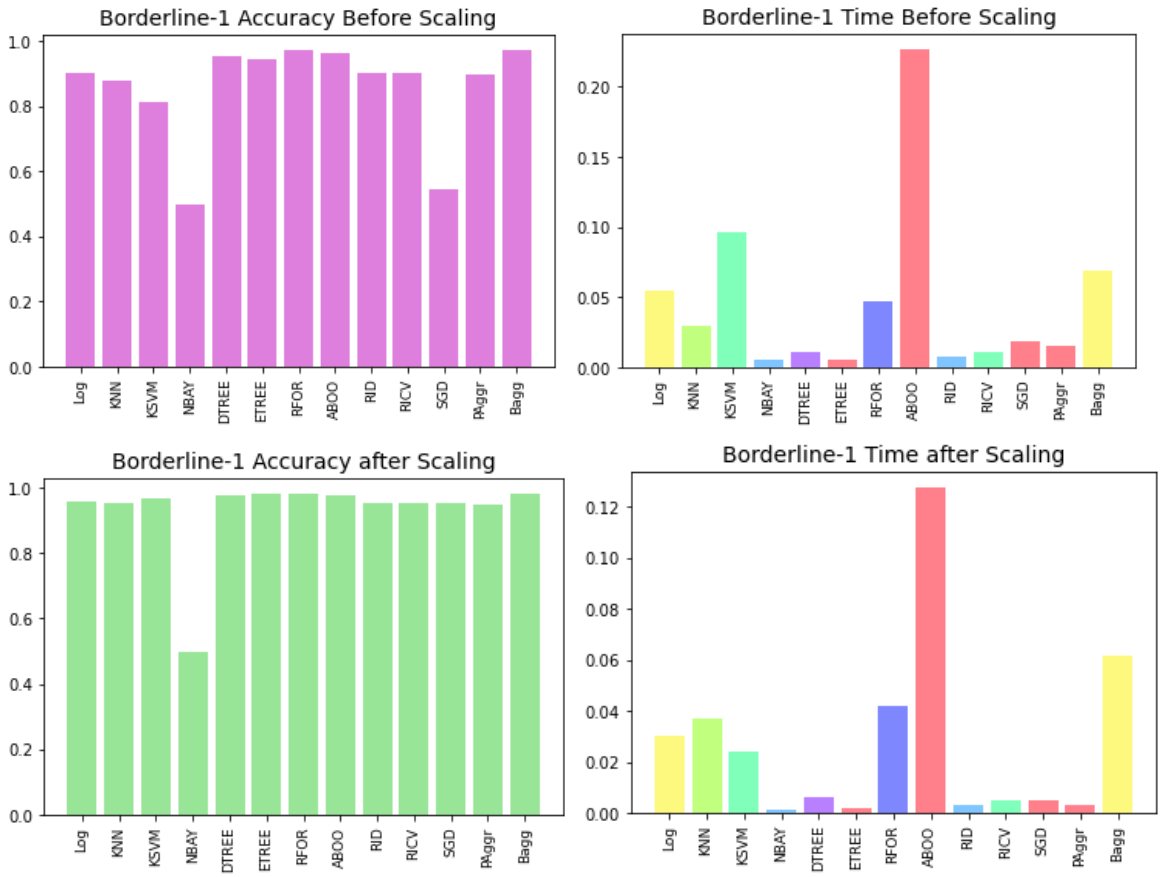

Figure 9 Border line 1 Oversampling Accuracy and Time Analysis pre and post scaling

Table 9

Borderline 1 oversampling metrics pre feature scaling

Classifier	Precision	Recall	FScore	Accuracy	Running Time (ms)
Logistic	0.90	0.90	0.90	0.90	0.05
KNN	0.89	0.88	0.88	0.88	0.03
KSVM	0.83	0.81	0.81	0.81	0.10
GNBayes	0.75	0.50	0.35	0.50	0.01
DTree	0.95	0.95	0.95	0.95	0.01
ETree	0.95	0.94	0.94	0.94	0.01
RForest	0.97	0.97	0.97	0.97	0.05
AdaBoost	0.96	0.96	0.96	0.96	0.23
Ridge	0.91	0.90	0.90	0.90	0.01
RidgeCV	0.91	0.90	0.90	0.90	0.01
SGD	0.77	0.55	0.44	0.55	0.02
PAggress	0.90	0.90	0.90	0.90	0.01
Bagging	0.97	0.97	0.97	0.97	0.07

Table 10

Borderline 1 oversampling metrics post feature scaling

Classifier	Precision	Recall	FScore	Accuracy	Running Time (ms)
Logistic	0.96	0.96	0.96	0.96	0.03
KNN	0.95	0.95	0.95	0.95	0.04
KSVM	0.97	0.97	0.97	0.97	0.02
GNBayes	0.75	0.50	0.34	0.50	0.00
DTree	0.98	0.98	0.98	0.98	0.01
ETree	0.98	0.98	0.98	0.98	0.00
RForest	0.98	0.98	0.98	0.98	0.04
AdaBoost	0.98	0.98	0.98	0.98	0.13
Ridge	0.95	0.95	0.95	0.95	0.00
RidgeCV	0.95	0.95	0.95	0.95	0.00
SGD	0.95	0.95	0.95	0.95	0.00
PAggress	0.95	0.95	0.95	0.95	0.00
Bagging	0.98	0.98	0.98	0.98	0.06

The unprocessed dataset is processed with Borderline 2 oversampling and then passed to all classifiers to examine the metrics and is shown in Figure. 10. The comparison of the efficiency metrics re and post scaling is shown in Table. 11 and Table. 12.

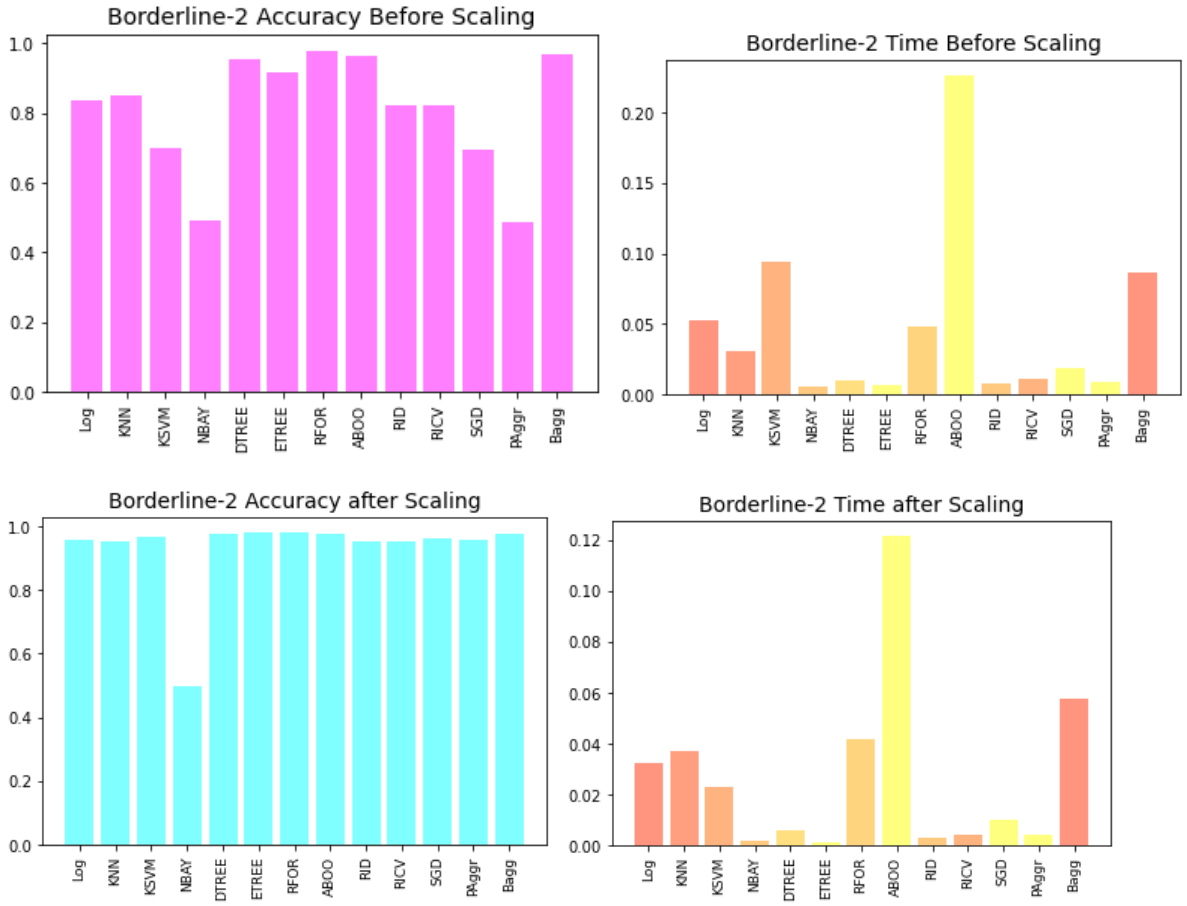

Figure 10. Borderline 2 Oversampling Accuracy and Time Analysis pre and post scaling

Table 11

Borderline 2 oversampling metrics pre feature scaling

Classifier	Precision	Recall	FScore	Accuracy	Running Time (ms)
Logistic	0.84	0.84	0.84	0.84	0.05
KNN	0.86	0.85	0.85	0.85	0.03
KSVM	0.72	0.70	0.69	0.70	0.09
GNBayes	0.61	0.49	0.35	0.49	0.01
DTree	0.95	0.95	0.95	0.95	0.01
ETree	0.92	0.92	0.92	0.92	0.01
RForest	0.98	0.98	0.98	0.98	0.05
AdaBoost	0.97	0.97	0.97	0.97	0.23
Ridge	0.83	0.82	0.82	0.82	0.01
RidgeCV	0.83	0.82	0.82	0.82	0.01
SGD	0.77	0.69	0.67	0.69	0.02
PAggress	0.23	0.48	0.32	0.48	0.01
Bagging	0.97	0.97	0.97	0.97	0.09

Table 12

Borderline 2 oversampling metrics post feature scaling

Classifier	Precision	Recall	FScore	Accuracy	Running Time (ms)
Logistic	0.96	0.96	0.96	0.96	0.03
KNN	0.95	0.95	0.95	0.95	0.04
KSVM	0.97	0.97	0.97	0.97	0.02
GNBayes	0.75	0.50	0.34	0.50	0.00
DTree	0.98	0.98	0.98	0.98	0.01
ETree	0.98	0.98	0.98	0.98	0.00
RForest	0.98	0.98	0.98	0.98	0.04
AdaBoost	0.98	0.98	0.98	0.98	0.12
Ridge	0.95	0.95	0.95	0.95	0.00
RidgeCV	0.95	0.95	0.95	0.95	0.00
SGD	0.96	0.96	0.96	0.96	0.01
PAggress	0.96	0.96	0.96	0.96	0.00
Bagging	0.98	0.98	0.98	0.98	0.06

Figure 10. Accuracy analysis of AllKNN dataset before and after feature scaling

The unprocessed dataset is processed with ADASYN oversampling and then passed to all classifiers to examine the metrics and is shown in Figure. 11. The comparison of the efficiency metrics re and post scaling is shown in Table. 13 and Table. 14.

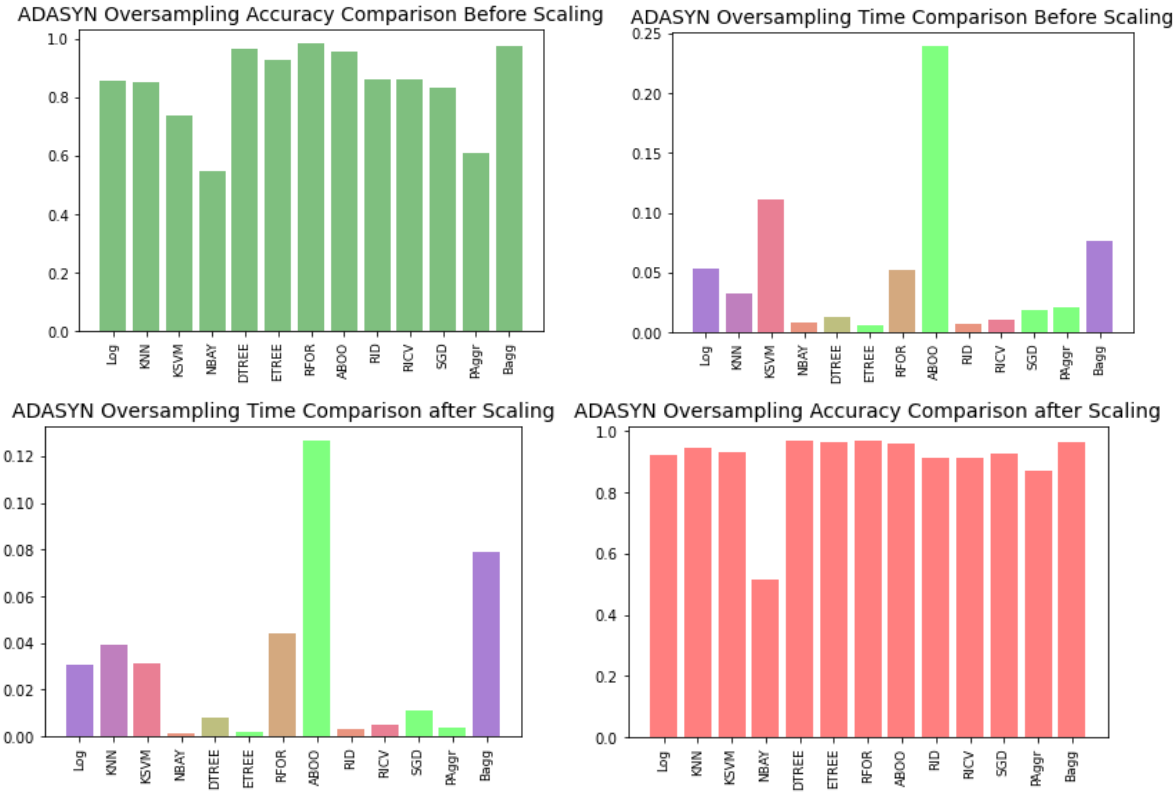

Figure 11. ADASYN Oversampling Accuracy and Time Analysis pre and post scaling

Table 13

ADASYN oversampling metrics pre feature scaling

Classifier	Precision	Recall	FScore	Accuracy	Running Time (ms)
Logistic	0.88	0.86	0.85	0.86	0.05
KNN	0.87	0.85	0.85	0.85	0.03
KSVM	0.78	0.74	0.73	0.74	0.11
GNBayes	0.76	0.55	0.40	0.55	0.01
DTree	0.96	0.96	0.96	0.96	0.01
ETree	0.93	0.93	0.93	0.93	0.01
RForest	0.98	0.98	0.98	0.98	0.05
AdaBoost	0.96	0.96	0.96	0.96	0.24
Ridge	0.88	0.86	0.86	0.86	0.01
RidgeCV	0.88	0.86	0.86	0.86	0.01
SGD	0.84	0.83	0.83	0.83	0.02
PAggress	0.77	0.61	0.52	0.61	0.02
Bagging	0.98	0.98	0.98	0.98	0.08

Table 14

ADASYN oversampling metrics post feature scaling

Classifier	Precision	Recall	FScore	Accuracy	Running Time (ms)
Logistic	0.92	0.92	0.92	0.92	0.03
KNN	0.95	0.94	0.94	0.94	0.04
KSVM	0.93	0.93	0.93	0.93	0.03
GNBayes	0.75	0.52	0.37	0.52	0.00
DTree	0.97	0.97	0.97	0.97	0.01
ETree	0.97	0.97	0.97	0.97	0.00
RForest	0.97	0.97	0.97	0.97	0.04
AdaBoost	0.96	0.96	0.96	0.96	0.13
Ridge	0.91	0.91	0.91	0.91	0.00
RidgeCV	0.91	0.91	0.91	0.91	0.00
SGD	0.93	0.93	0.93	0.93	0.01
PAggress	0.88	0.87	0.87	0.87	0.00
Bagging	0.97	0.97	0.97	0.97	0.08

Conclusion

This work sought to analyze the implementation of non-sampled target characteristics using tested data. The target biospy is found be non-sampled with 93.6% of health and 6.4% of cervical cancer patients. So this workflow aims to oversample the healthy and cervical cancer classes to be equalized to 93.6% inorder to improvise the accuracy of the cervical cancer prediction. The unprocessed dataset is applied with all the oversampling methods and the

oversampled dataset is executed with all the classifiers to examine the performance metrics along with the execution time. The findings of the experiments reveal that the Random forest classifier tends to sustain 96% accuracy pre and post scaling for unporocessed dataset. Similarly the same classifier tends to sustain 98% accuracy for all the oversampling techniques.

References

1. Adedimeji, A., Ajeh, R., Pierz, A. *et al.* (2021). Challenges and opportunities associated with cervical cancer screening programs in a low income, high HIV prevalence context. *BMC Women's Health,* 21(74).
2. Nazim Razali., Salama Mostafa, A., Aida Mustapha., Mohd Helmy., Abd Wahab., Nurul Atieqah Ibrahim. (2020). Risk Factors of Cervical Cancer using Classification in Data Mining. Journal of Physics: Conference Series, (1529), pp 25-27.
3. Alam., Talha Mahboob Khan., Muhammad Milhan Afzal Iqbal., Muhammad Atif Abdul., Wahab Mushtaq., Mubbashar. (2019) Cervical Cancer Prediction through Different Screening Methods Using Data Mining. *International Journal of Advanced Computer Science and Applications*, 10(2).
4. Subramanian, S., Sankaranarayanan, R., Esmy, P, O., Thulaseedharan, J, V., Swaminathan, R., Thomas, S. (2016). Clinical trial to implementation: Cost and effectiveness considerations for scaling up cervical cancer screening in low-and middle-income countries. *Journal of Cancer Policy*, 7, pp. 4 – 11
5. Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R, L., Torre, L, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 68, pp. 394 – 424
6. Ahmad, P., Qamar, S., Rizvi, S, Q, A. (2015) Techniques of data mining in healthcare: a review. *International Journal of Computer Applications*, 120.
7. Chang, C C., Cheng, S L., Lu, C, J., Liao, K H. (2013). Prediction of Recurrence in Patients with Cervical Cancer Using MARS and Classification. *International Journal of Machine Learning and Computing*, 3

8. Fernandes, K., Chicco, D., Cardoso, J S., Fernandes, J. (2018). Supervised deep learning embeddings for prediction of cervical cancer diagnosis. *PeerJ Computer Science*, 4.

9. Chicco, D., Masseroli, M. (2015). Software suite for gene and protein annotation prediction and similarity search. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 12(4), pp. 837-843.

10. Kauffman, R, P., Griffin, S, J., Lund, J, D., Tullar, P, E. (2013). Current recommendations for cervical cancer screening: do they render the annual pelvic examination obsolete. *Medical Principles and Practice*, 22(4), pp. 313-322.

11. Vidya, R., Nasira, G. (2016). Prediction of cervical cancer using hybrid induction technique: solution for human hereditary disease patterns. *Indian Journal of Science and Technology*, 9.

12. Karthick, R., and M. Sundararajan. "SPIDER-based out-of-order execution scheme for HtMPSOC." International Journal of Advanced Intelligence paradigms 19.1 (2021): 28-41. https://doi.org/10.1504/IJAIP.2021.114581

13. Sabarish, P., et al. "An Energy Efficient Microwave Based Wireless Solar Power Transmission System." IOP Conference Series: Materials Science and Engineering. Vol. 937.No. 1.IOP Publishing, 2020. doi:10.1088/1757-899X/937/1/012013

14. Vijayalakshmi, S., et al. "Implementation of a new Bi-Directional Switch multilevel Inverter for the reduction of harmonics." IOP Conference Series: Materials Science and Engineering. Vol. 937.No. 1.IOP Publishing, 2020. doi:10.1088/1757-899X/937/1/012026

15. P. Sabarish, R. Karthick, A. Sindhu, N. Sathiyanathan, Investigation on performance of solar photovoltaic fed hybrid semi impedance source converters, Materials Today: Proceedings, 2020, https://doi.org/10.1016/j.matpr.2020.08.390

16. Karthick, R., and M. Sundararajan. "A novel 3-D-IC test architecture-a review." International Journal of Engineering and Technology (UAE) 7.1.1 (2018): 582-586.

17. Karthick, R., and M. Sundararajan. "Design and implementation of low power testing using advanced razor based processor." International Journal of Applied Engineering Research 12.17 (2017): 6384-6390.

18. Suresh, HelinaRajini, et al. "Suppression of four wave mixing effect in DWDM system." Materials Today: Proceedings (2021). https://doi.org/10.1016/j.matpr.2020.11.545

19. M Ramkumar, C Ganesh Babu, K Vinoth Kumar, D Hepsiba, A Manjunathan, R Sarath Kumar, "ECG Cardiac arrhythmias Classification using DWT, ICA and MLP Neural Network", Journal of Physics: Conference Series, vol.1831, issue.1, pp.012015, 2021
20. M Ramkumar, C Ganesh Babu, A Manjunathan, S Udhayanan, M Mathankumar,"A Graphical User Interface Based Heart Rate Monitoring Process and Detection of PQRST Peaks from ECG Signal" Lecture Notes in Networks and Systems, 2021, 173 LNNS, pp. 481–496
21. A Manjunathan, A Lakshmi, S Ananthi, A Ramachandran, C Bhuvaneshwari, "Image Processing Based Classification of Energy Sources in Eatables Using Artificial Intelligence", Annals of the Romanian Society for Cell Biology,vol.25, issue.3, pp.7401-7407, 2021.