“Imbalanced dataset classification by using oversampling and under-sampling techniques’’

Amna Qudrat[1], M. Hamza Kabeer [2]
amnaqudrat@outlook.com
hamzakabeercs@gamil.com

Department of Computer Science, Lahore Garrison University, DHA phase 6, Lahore, Pakistan

***Abstract~~* Imbalanced data set problem occurs in classification, where the number of cases of one class is much inferior than the cases of the other classes. The main task in imbalance problem is that the irrelevant classes are often more beneficial, but normal classifiers tend to be weighed down by the vast classes and overlook the small ones. In machine learning the imbalanced datasets has become a serious problem and to solve this problem we will use two famous techniques over-sampling and under-sampling and investigate which method scores best. Therefore, in this paper, we present an experiential study about the use of over-sampling and under-sampling methods to advance the accuracy of case selection methods on imbalanced datasets.**

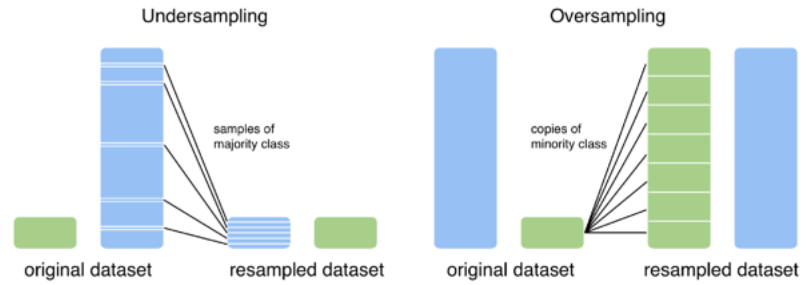
Keywords: imbalanced data, resampling, oversampling, undersampling. Machine learning, tomek links

I. Introduction:

Imbalanced dataset classification with Over-sampling and under-sampling techniques and Investigate which of the methods score best. Over-sampling and under-sampling in data analysis are techniques used to adjust the class distribution of a data set (i.e. the ratio between the different classes/categories represented). These terms are used both in statistical sampling, survey design methodology and in machine learning.

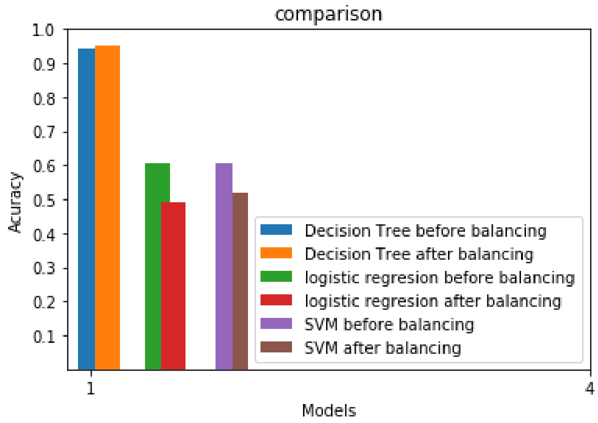
We Balanced/Resamples a dataset by applying the Synthetic Minority Oversampling Technique along with Tomek links technique that remove unwanted overlap between classes where majority class links are removed until all minimally distanced nearest neighbor pairs are of the same class. We apply different oversampling and under-sampling methods jointly with instance selectors over several public imbalanced databases. Our experimental results show that using oversampling and under-sampling methods significantly improves the accuracy for the minority class.

In this work, the imbalanced dataset of breast cancer patients and dataset from Kaggle has been used with different machine learning models to experiment over-sampling and under-sampling techniques and apply a full comparison with different appraisal metrics. The outcomes show that over-sampling technique has well scores than the under-sampling technique for different machine learning classifier models. This paper is systematized as follows: related work is shown in section II. Section III describes the dataset that has been used in this article. Our methodology and evaluation metrics are presented in section IV. Experiments and results are introduced in section V. Finally, the conclusion of the paper is provided in section VI.



II. RELATED WORK:

The imbalance data challenge has concerned increasing devotion of researchers, recently. Authors D. L. Wilson proposed a renowned technique for under-sampling. It works by removing the data points where goal class does not equivalent the majority of its KNN. In 2002, the two authors discussed several problems related to Learning with skewed class distributions. For example, the connection between class scatterings and price sensitive knowledge, and the boundaries of error frequency and accuracy to measure the act of models. In 2005, some other authors name S. Visa and A. Ralescu, proposed a review for the most commonly used methods Issues in mining imbalanced data sets-a review paper. They claimed that the bad performance of the models created by the typical machine datasets has no missing values.



The figure shows the result which we apply on imbalanced dataset of breast cancers cell classification and get the result before and after balancing. The results are more accurate after applying over-sampling and under-sampling balancing techniques.

III. DATA BASE:

Data description: this breast cancer databas was obtained from the university of Wiscinsin Hospital, Madison from Dr. William H. Wollberg. please see: O.L.Mangasarian and W. H. Wolberg: "Cancer diagnosis via linear programming",SIAM News, Volume 23, Number 5, September 1990, pp. 1 to 18.

Information:

- Number of Cases: 699
- The 16 Cases with missing attribute values are unconcerned from the database, leaving 683 cases.
- Number of attributes: 10 plus the class attribute
- Attribute 2 through 10 will be used to represent cases.
- Each case has one of 2 possible classes: benign or malignant.
- Class distribution:
- Benign: 458(65.5%)
- Malignant: 241 (34.5%)

Attribute Information:

Attribute	Domain
1.Sample code number (id number)	1-10
2.Clump Thickness	1-10
3.Uniformity of Cell Size	1-10
4.Uniformity of Cell Shape	1-10
5.Marginal Adhesion	1-10
6.Single Epithelial Cell Size	1-10
7.Bare Nuclei	1-10
8.Bland Chromatin	1-10
9.Normal Nucleoli	1-10
10.Mitoses	1-10
11.Class (2 for benign, 4 for malignant)	1-10

The cancer database file has 663 cases. Since this database does not have enough cases, 100% accuracy will not be accomplished in this circumstance. However, even with this restricted database, the DecisionMaker will be able to give an 85% general accuracy rate. The number of mandatory cases to achieve 100% accuracy and the number of existing cases in this situation is listed below.

Class	Required Cases	Actual Cases
2=benign	1000	458
4=malignant	1000	241
Total	2000	663

IV. Methodology:

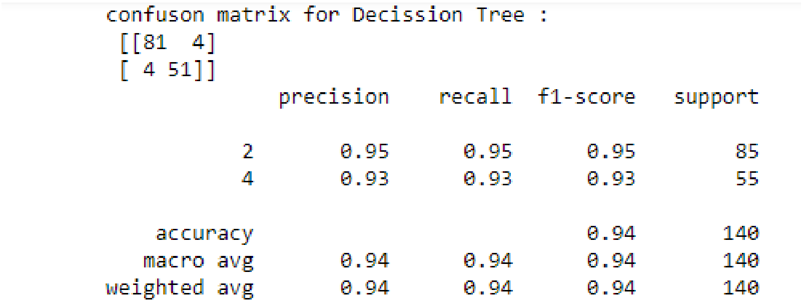
In order to give a complete understanding of the unbalanced data problem, we have started with discovering the dataset. After taking the imbalanced dataset of breast cancer patients and dataset from Kaggle, we splinted dataset into testing and training dataset. we applied Decision Tree, Logistic Regression, SVM before balancing, Gaussian NB before balancing and compared/tested accuracy score and studied confusion matrix for each technique then we applied Tomeklink+SMOTE combine technique for balancing our dataset. Then using K-fold cross validation we compared Accuracy of each above-mentioned techniques.

Abbreviation	**Machine Learning classifier models**
SVM(Linear)	Support Vector Machine with Linear kernel
SVM(Poly)	Support Vector Machine with Poly kernel
SVM(RBF)	Support Vector Machine with RBF kernel
NB	Gaussian Naive Bayes
LR	Logistic regression
DT1	Decision tree
DT2	Decision tree
DT3	Decision tree
RF	Random Forest
GB	Gradient Boosting Classifier

V. Experiments and results:

When we apply sampling techniques on our given imbalanced dataset. The different results are shown before and after applying methods.

1. confuson matrix for Decission Tree shows the result in this form,

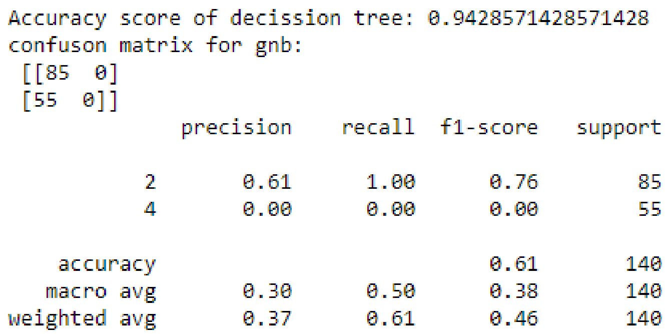


```
confuson matrix for Decission Tree :
 [[81  4]
 [ 4 51]]
              precision    recall  f1-score   support

           2       0.95      0.95      0.95        85
           4       0.93      0.93      0.93        55

    accuracy                           0.94       140
   macro avg       0.94      0.94      0.94       140
weighted avg       0.94      0.94      0.94       140
```

2. Accuracy score of decission tree:

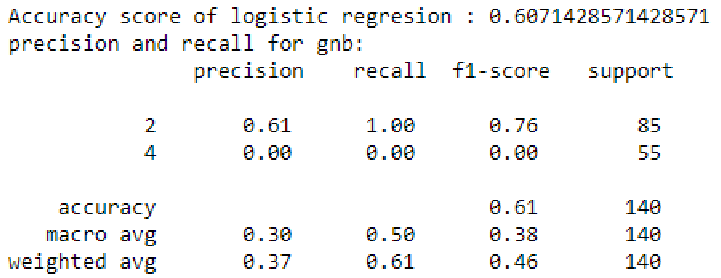


```
Accuracy score of decission tree: 0.9428571428571428
confuson matrix for gnb:
 [[85  0]
 [55  0]]
              precision    recall  f1-score   support

           2       0.61      1.00      0.76        85
           4       0.00      0.00      0.00        55

    accuracy                           0.61       140
   macro avg       0.30      0.50      0.38       140
weighted avg       0.37      0.61      0.46       140
```

3. Accuracy score of logistic regression:



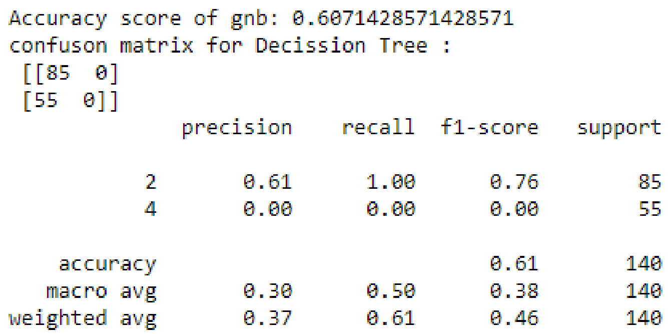
```
Accuracy score of logistic regresion : 0.6071428571428571
precision and recall for gnb:
              precision    recall  f1-score   support

           2       0.61      1.00      0.76        85
           4       0.00      0.00      0.00        55

    accuracy                           0.61       140
   macro avg       0.30      0.50      0.38       140
weighted avg       0.37      0.61      0.46       140
```

When check accuracy by using logistic regression on the basics of   precision and recall for gnb.

4. Accuracy score of gnb:

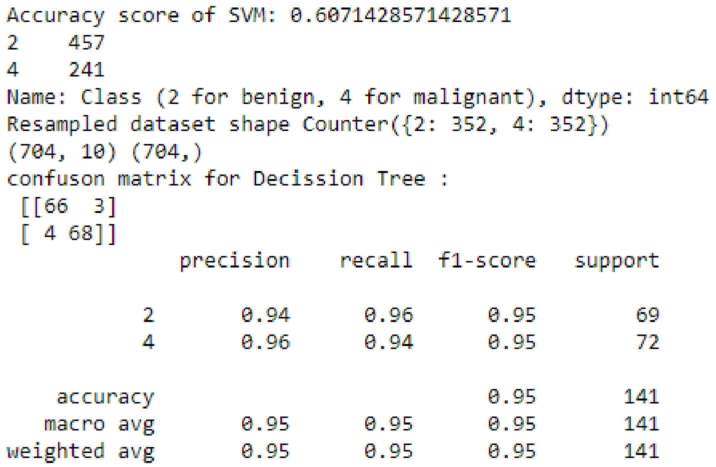


```
Accuracy score of gnb: 0.6071428571428571
confuson matrix for Decission Tree :
 [[85  0]
 [55  0]]
              precision    recall  f1-score   support

           2       0.61      1.00      0.76        85
           4       0.00      0.00      0.00        55

    accuracy                           0.61       140
   macro avg       0.30      0.50      0.38       140
weighted avg       0.37      0.61      0.46       140
```

5. Accuracy score of SVM:



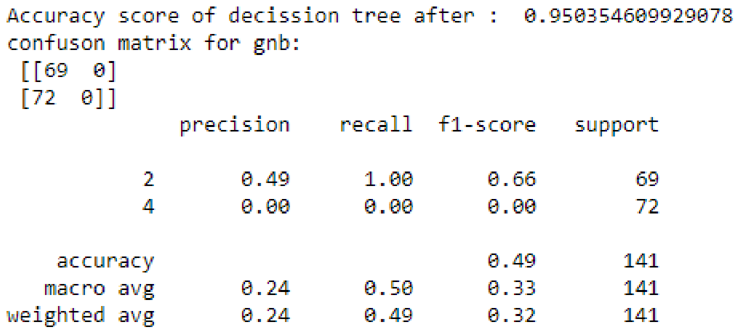
```
Accuracy score of SVM: 0.6071428571428571
2    457
4    241
Name: Class (2 for benign, 4 for malignant), dtype: int64
Resampled dataset shape Counter({2: 352, 4: 352})
(704, 10) (704,)
confuson matrix for Decission Tree :
 [[66  3]
 [ 4 68]]
              precision    recall  f1-score   support

           2       0.94      0.96      0.95        69
           4       0.96      0.94      0.95        72

    accuracy                           0.95       141
   macro avg       0.95      0.95      0.95       141
weighted avg       0.95      0.95      0.95       141
```

Now the accuracies after applying sampling techniques.

1. Accuracy score of decision tree after:

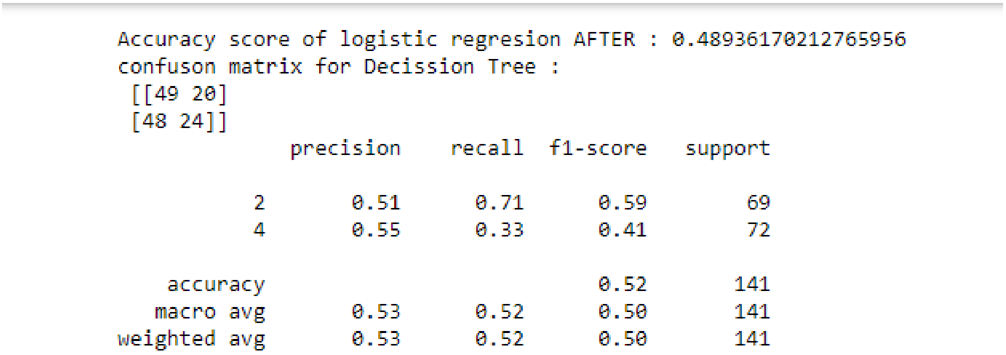


```
Accuracy score of decission tree after :  0.950354609929078
confuson matrix for gnb:
 [[69  0]
 [72  0]]
              precision    recall  f1-score   support

           2       0.49      1.00      0.66        69
           4       0.00      0.00      0.00        72

    accuracy                           0.49       141
   macro avg       0.24      0.50      0.33       141
weighted avg       0.24      0.49      0.32       141
```

2. Accuracy score of logistic regression AFTER:

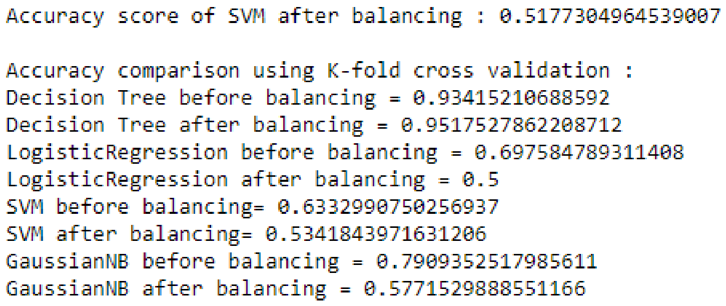


```
Accuracy score of logistic regresion AFTER : 0.48936170212765956
confuson matrix for Decission Tree :
 [[49 20]
 [48 24]]
              precision    recall  f1-score   support

           2       0.51      0.71      0.59        69
           4       0.55      0.33      0.41        72

    accuracy                           0.52       141
   macro avg       0.53      0.52      0.50       141
weighted avg       0.53      0.52      0.50       141
```

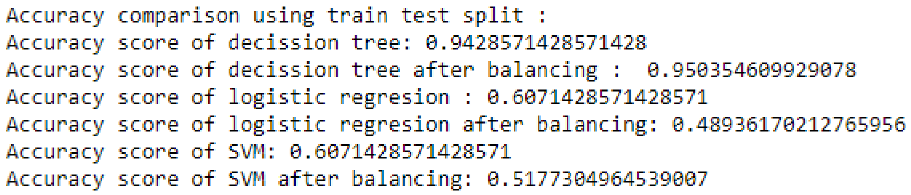
3. Accuracy score of SVM after balancing:



```
Accuracy score of SVM after balancing : 0.5177304964539007

Accuracy comparison using K-fold cross validation :
Decision Tree before balancing = 0.93415210688592
Decision Tree after balancing = 0.9517527862208712
LogisticRegression before balancing = 0.697584789311408
LogisticRegression after balancing = 0.5
SVM before balancing= 0.6332990750256937
SVM after balancing= 0.5341843971631206
GaussianNB before balancing = 0.7909352517985611
GaussianNB after balancing = 0.5771529888551166
```

4. Accuracy comparison using train test split:



```
Accuracy comparison using train test split :
Accuracy score of decission tree: 0.9428571428571428
Accuracy score of decission tree after balancing :  0.950354609929078
Accuracy score of logistic regresion : 0.6071428571428571
Accuracy score of logistic regresion after balancing: 0.48936170212765956
Accuracy score of SVM: 0.6071428571428571
Accuracy score of SVM after balancing: 0.5177304964539007
```

VI. Conclusion:

- By studying graphs plotted for each technique on unbalanced dataset and balanced dataset we can conclude that the given data set for studying tumor cell the Decision tree algorithm after performing dataset balancing the results are more accurate and are reliable.
- we are not liable that these techniques will shows always accurated results using different datasets of different fields.

References.

[1] H. He and E. A. Garcia, “Learning from imbalanced data,” IEEE Transactions on knowledge and data engineering, vol. 21, no. 9, pp. 1263–1284, 2009.

[2] I. Tomek, “A generalization of the k-nn rule,” IEEE Transactions on Systems, Man, and Cybernetics, no. 2, pp. 121–126, 1976.

[3] G. Lemaˆıtre, F. Nogueira, and C. K. Aridas, “Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning,” The Journal of Machine Learning Research, vol. 18, no. 1, pp. 559–563, 2017.

[4] H. Han, W.-Y. Wang, and B.-H. Mao, “Borderline-smote: a new oversampling method in imbalanced data sets learning,” in International conference on intelligent computing. Springer, 2005, pp. 878–887.

[5] https://github.com/Roweida-Mohammed/Code For Santander Customer Transaction Prediction.

[6] D. L. Wilson, “Asymptotic properties of nearest neighbor rules using edited data,” IEEE Transactions on Systems, Man, and Cybernetics, no. 3, pp. 408–421, 1972.

[7] M. C. Monard and G. E. Batista, “Learmng with skewed class distrihutions,” Advances in Logic, Artificial Intelligence, and Robotics: LAPTEC, vol. 85, no. 2002, p. 173, 2002.

[8] S. Visa and A. Ralescu, “Issues in mining imbalanced data sets-a review paper,” in Proceedings of the sixteen midwest artificial intelligence and

cognitive science conference, vol. 2005. sn, 2005, pp. 67–73.
[9] Julio Hernandez, Jes´us Ariel Carrasco-Ochoa,
and Jos´e Francisco Mart´ınez-Trinidad
Instituto Nacional de Astrof´ısica Optica y Electr´ ´ onica, Computer Science
Department, Luis Enrique Erro No. 1, Sta. Mar´ıa Tonantzintla,
Puebla, CP 72840, Mexico
{julio.hernandez.t,ariel,fmartine}@ccc.inaoep.mx
http://ccc.inaoep.mx
[10] H. Rathpisey and T. B. Adji, "Handling Imbalance Issue in Hate Speech Classification using Sampling-based Methods," 2019 5th International Conference on Science in Information Technology (ICSITech), Yogyakarta, Indonesia, 2019, pp. 193-198, doi: 10.1109/ICSITech46713.2019.8987500.