Avrupa Bilim ve Teknoloji Dergisi
Özel Sayı 34, S. 142-147, Mart 2022
© Telif hakkı EJOSAT'a aittir
Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
Special Issue 34, pp. 142-147, March 2022
Copyright © 2022 EJOSAT
Research Article

Evaluation of Oversampling Methods (OVER, SMOTE, and ROSE) in Classifying Soil Liquefaction Dataset based on SVM, RF, and Naïve Bayes

Selçuk Demir[1*], Emrehan Kutluğ Şahin[2]

[1*] Bolu Abant Izzet Baysal University, Departmant of Civil Engineering, Bolu, Turkey, (ORCID: 0000-0003-2520-4395), selcukdemir@ibu.edu.tr
[2] Bolu Abant Izzet Baysal University, Departmant of Civil Engineering, Bolu, Turkey, (ORCID: 0000-0002-9830-8585), emrehansahin@ibu.edu.tr

(2nd International Conference on Applied Engineering and Natural Sciences ICAENS 2022, March 10-13, 2022)

(**DOI:** 10.31590/ejosat.1077867)

ATIF/REFERENCE: Demir, S. & Sahin, E. (2022). Evaluation of oversampling methods (OVER, SMOTE, and ROSE) in classifying soil liquefaction dataset based on SVM, RF, and Naïve Bayes. *European Journal of Science and Technology*, (34), 142-147.

Abstract

Class imbalanced datasets are prevalent in real-world applications, including engineering, medical domain, financial sector, and others. Machine learning (ML)-based prediction models have successfully demonstrated the applicability of various algorithms for the solution of different problems. However, their application for the soil liquefaction issue considering the class imbalance situation is limited. This paper presents the prediction results of random forest (RF), support vector machine (SVM), and naïve bayes (NB) algorithms with different training sample sizes for soil liquefaction. The effect of oversampling methods, namely simple oversampling (OVER), random oversampling examples (ROSE), and synthetic minority oversampling technique (SMOTE), on the prediction performance of classification algorithms is also investigated. Performance results are evaluated by means of some metrics, including Accuracy, Kappa, Precision, Recall, and F-measure. The results concluded the effectiveness of applying oversampling methods on imbalanced data before the modeling phase. All of the oversampling methods helped to enhance the overall performances of the classification models. It is also observed that the SMOTE exhibited slightly better performance than other considered oversampling methods. Furthermore, the SVM model outperformed compared to RF and NB models when all algorithms were trained by the SMOTE algorithm.

Keywords: Liquefaction Prediction, Naïve Bayes, Imbalanced Data, RF, SVM, Oversampling

SVM, RF ve Naive Bayes'e Dayalı Olarak Zemin Sıvılaşma Veri Setinin Sınıflandırılmasında Aşırı Örnekleme Yöntemlerinin (OVER, SMOTE ve ROSE) Değerlendirilmesi

Öz

Dengesiz sınıf veri kümeleri, mühendislik, tıp alanı, finans sektörü ve diğerleri dahil olmak üzere gerçek dünya uygulamalarında oldukça yaygındır. Makine öğrenimi (ML) tabanlı tahmin modelleri, farklı problemlerin çözümü için çeşitli algoritmaların uygulanabilirliğini başarıyla göstermiştir. Ancak sınıf dengesizliği durumu göz önüne alındığında zemin sıvılaşması sorununa yönelik uygulamaları sınırlıdır. Bu çalışma, zemin sıvılaşması için farklı eğitim örneği boyutlarına sahip rastgele orman (RF), destek vektör makinesi (SVM) ve naive bayes (NB) algoritmalarının tahmin sonuçlarını sunmaktadır. Ayrıca, basit aşırı örnekleme (OVER), rastgele aşırı örnekleme örnekleri (ROSE) ve sentetik azınlık aşırı örnekleme tekniğinin (SMOTE) gibi aşırı örnekleme yöntemlerinin sınıflandırma algoritmalarının tahmin performansı üzerindeki etkisi araştırılmıştır. Performans sonuçları, Accuracy, Kappa, Precision, Recall ve F-measure gibi metrikler aracılığıyla değerlendirilmiştir. Sonuçlar, modelleme aşamasından önce dengesiz veriler üzerinde aşırı örnekleme yöntemlerinin uygulanmasının etkili olduğu göstermiştir. Ayrıca, bütün aşırı örnekleme yöntemlerinin, sınıflandırma modellerinin genel performanslarını geliştirmeye yardımcı olduğu görülmüştür. SMOTE yönteminin diğer dikkate alınan aşırı örnekleme yöntemlerinden biraz daha iyi performans gösterdiği gözlemlenmiştir. Bununla beraber, bütün algoritmalar SMOTE algoritması ile eğitildiğinde, SVM modeli RF ve NB modellerine kıyasla daha iyi performans sergilemiştir.

Anahtar Kelimeler: Sıvılaşma Tahmini, Naïve Bayes, Dengesiz Veri Seti, RF, SVM, Aşırı Örnekleme

[*] Corresponding Author: selcukdemir@ibu.edu.tr

1. Introduction

Natural disasters (i.e., earthquakes, flood, hurricanes, volcanic eruptions, and others) are an ever-present danger to modern societies throughout the world. Among natural disasters, earthquakes stand out in seismic-prone areas due to their unpredictable catastrophic effects on societies and economies. Earthquakes appear without warning and devastate a region within seconds, thereby causing environmental damages, loss of lives, social and economic breaks. Besides, different types of earthquake-induced effects, such as landslides, liquefaction, and tsunami may be observed in zones of high seismicity. The seismic soil liquefaction is main responsible for the devastating hazards because of its mechanical structure. Liquefaction is commonly observed in soil deposits that are loose, cohesionless, and fine grained with high groundwater levels (Allen, 1982). During the liquefaction phenomenon, solid granular materials transform a liquefied state due to the porewater pressure-related loss of soil stiffness and shear strength induced by strong earthquakes. Thus, different types of failures associated with liquefiable soils have been observed from post-earthquake observations, such as excessive settlements, lateral spreads, ground cracks, and sand boils (Adalıer and Elgamal, 2004). Due to the liquefaction-induced damages caused by major earthquakes, researchers and engineers have attempted to establish a better understanding of liquefaction and its effects for appropriately designing engineering structures against soil liquefaction.

Several liquefaction evaluation procedures (e.g., Cetin et al., 2004; Robertson and Wride, 1998; Kayen et al., 2013) have been proposed through standard penetration test (SPT), cone penetration test (CPT), and shear wave velocity test (Vs). However, among these procedures, the SPT-based simplified procedure (Seed and Idriss, 1971) is the first and simplest methods used for evaluating the seismic liquefaction resistance of soils in the field of engineering. Besides, experimental and numerical studies are other essential ways used for understanding the mechanism and influencing factors of liquefaction. On the other hand, in recent years, various machine learning (ML) algorithms have been proposed for engineering problems with the development of soft-computing tools. These algorithms provide a powerful tool to solve most problems having high complexity.

ML-based applications have made great progress in geotechnical engineering. Different kinds of ML tools, including artificial neural network (ANN), random forest (RF), support vector machine (SVM), eXtreme gradient boosting (XGBoost), canonical correlation forest (CCF), k-nearest neighbors (kNN), deep neural network (DNN), etc. have been successfully employed in several geotechnical applications (Koopialipoor et al., 2020; Demir and Sahin, 2022; Samui, 2008; Wang et al., 2020; Amiri et al., 2016; Zhang et al., 2021a). The use of ML algorithms in liquefaction issues has also considered for classifying soil liquefaction or predicting the liquefaction-induced lateral spreads by the use of regression procedures (Xie et al., 2020). Some of the recent studies for liquefaction prediction are briefly mentioned here. For example, Zhang et al. (2021b) employed the DNN strategy to predict soil liquefaction based on the Vs and SPT dataset. Zhou et al. (2021) proposed two support vector machine (SVM) models for predicting liquefaction potential using genetic algorithm (GA) and grey wolf optimizer (GWO) techniques in order to enhance the efficiency of the models. Zhao et al. (2021) developed the kernel extreme learning machine (KELM) with particle swarm optimization (PSO) based soil liquefaction potential evaluation system using CPT and Vs measurements. Hu et al. (2021) used Bayesian network (BN) model for soil liquefaction prediction under the conditions of nine different training sample size ratios. Demir and Sahin (2022) investigated the performance of three forest algorithms to predict the liquefaction potential of soils from two different CPT datasets using CCF, RF, and rotation forest (RotFor). These studies revealed that ML algorithms provide feasible solutions to tackle soil liquefaction prediction problems. Nevertheless, these studies have tended to focus on some factors, such as the applicability of ML algorithms, the effects of optimization approaches on ML algorithms, and ratios of training sample size.

There is an important topic in ML that should be kept in mind when working with traditional classifiers, and that is class imbalance. In the case of class imbalance, the distribution of classes in a dataset is one-sided that means the number of samples in some classes is more than other classes. Generally, the class imbalance ratio (IR), defined as the number of samples in the majority class divided by the number of samples in the minority class, is used to show the degree of imbalance of a dataset. Any dataset with an IR value is close to or exceeding 1.5-2.0 is considered imbalanced (He and Ma, 2013; Vluymans, 2018). Traditional classification algorithms result poorly in imbalanced datasets because they are designed for balanced datasets (Douzas and Bacao, 2020). The target to be reached in these algorithms is to perform the best prediction accuracy by adjusting loss functions to minimize the losses, which is biased to the majority class (Chen et al., 2021). For that reason, handling an imbalanced dataset is crucial for the successful development of a prediction model. Several techniques exist to solve class imbalance problem. These techniques can be divided into four main categories, depending on how they deal with the problem: algorithm-level methods, data-level methods, cost-sensitive methods, and ensemble-based methods (Fernández et al., 2018). Among them, data-level methods are standard techniques in imbalanced learning, they are widely used in data science problems. The data-level methods, which are categorized into three groups, oversampling, undersampling, and hybrid methods, aim to change the class distribution by manipulating the training data towards a more balanced one. Since undersampling methods eliminate data in the majority class, which causes the loss of important data, oversampling is more frequently preferred than the other data-level methods.

In this study, the performances of three popular classification methods, RF, SVM, and Naïve Bayes (NB) coupled with three oversampling methods, namely simple oversampling (OVER), random oversampling examples (ROSE), and synthetic minority oversampling technique (SMOTE) were analyzed using a CPT-based liquefaction dataset. The effect of three oversampling methods was also compared with respect to a well-known sampling method called simple random sampling (SRS). Finally, performance metrics of RF, SVM, and NB in prediction soil liquefaction were presented using the confusion matrix.

2. Material and Method

2.1. Introduction of the Dataset

The dataset used in this study is based on historical CPT case records taken from six different earthquakes (4 in the U.S. and 2 in China and Taiwan) reported by Juang et al. (2003) This dataset includes 226 cases, 133 non-liquefied (No) and 93 liquefied (Yes) cases. A total of seven features, i.e., depth of the soil layer (*d*, m), cone tip resistance (q_c, MPa), the sleeve friction ratio (R_f, %), the total and effective vertical stresses (σ_v and σ'_v, kPa), the peak ground acceleration (a_{max}, g), and the earthquake magnitude moment (M_w) were used as the input parameters. The statistical ranges of values associated with each input parameter are given in Table 1. In order to gain a better insight into the relationship between the input variables, a scatter plot matrix was plotted as shown in Fig. 1. This matrix depicts the distributions of the variables, their correlation coefficients between each other as well as their individual histogram plots.

Table 1. Some statistical measures of the used dataset

Variable	Min-Max		Mean	Sd	Median
d	1.4	16.5	5.67	2.93	4.8
q_c	0.9	25	5.82	4.09	4.9
R_f	0.1	5.2	1.22	1.05	0.9
σ'_v	22.5	215.2	74.65	34.4	62.8
σ_v	26.6	274	106.89	55.36	90.3
a_{max}	0.08	0.8	0.29	0.14	0.25
M_w	6	7.6	6.95	0.44	7.1

2.2. Classification Methods

2.2.1. Naïve Bayes

Naïve Bayes (NB) algorithm is a probabilistic model using Bayes' theorem, which requiring a small amount of training data to estimate the statistical parameters (such as mean and variance) necessary for the classification. NB considers the strong or naive independence of data points. The NB classifiers are used for many classification problems (e.g., text analysis, document classification, signal segmentation, natural hazards, and medical diagnosis). Due to being simple to implement, this classifier is a preferred method in ML, and this is an important advantage of NB to the other ML methods for classification purposes.

2.2.2. Random Forest

Random Forest (RF) is an ensemble classifier method based on bagging and decision trees (DTs) that uses multiple models of various DTs to improve prediction accuracy. This algorithm is well suitable for handling classification and regression problems. The main concept of RF is to independently build multiple DTs by bootstrap samples from the original training dataset. The averages of the predictions of these single trees are used to obtain an accurate and stable prediction (He et al., 2022). RF creates numerous trees, which limits generalization error because of the ensemble of permutations that can cope with the classification error of one permutation. Thus, the RF method can ensure great improvements in classification accuracy and can be easily implemented for parallel computing which makes it a popular choice for data classification and computationally efficient (Wu et al., 2020).

2.2.3. Support Vector Machine

Support Vector Machine (SVM) is a very powerful supervised ML algorithm used to handle a two-class pattern recognition problem for classification and regression analysis. SVMs have been applied successfully to many engineering related applications in recent years. In SVM, the data is evaluated, and patterns are identified in order to create a classification. The goal here is to identify the best optimal hyperplane between two classes by maximizing the margin between their nearest points. The effectiveness of SVM is determined by the kernel type and parameters. SVM algorithms use differing kinds of kernel functions namely linear, nonlinear, polynomial, radial basis function, and sigmoid. In the present work, radial basis function was used for the SVM models as kernel functions because of its efficiency in providing very high prediction performance. Before estimating the model, SVM with radial basis function needs to tune two hyperparameters (e.g., *C* and *gamma*).

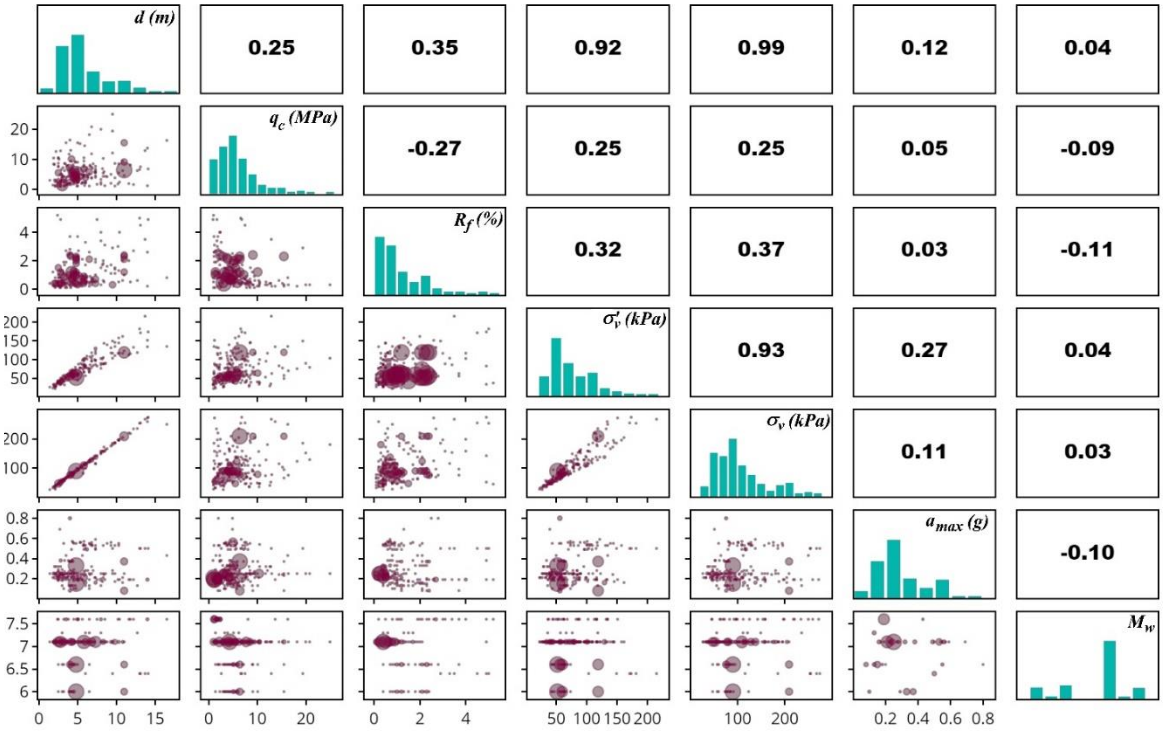

Figure 1. Scatter plot, histogram, and Pearson correlation coefficients of the input parameters

2.3. Oversampling Methods for Imbalanced Dataset

The aim of the resampling is to balance the data, which means that the ratio of the majority/minority class should be close to 1.0. The resampling methods can be divided in three groups namely undersampling, oversampling, and hybrid. Undersampling means to reduce the size of the majority class down to that of the minority class. Undersampling is a good choice for fasting computer processes, but it may not produce a decent model because it loses information by eliminating a fraction of the majority observations. In the case of oversampling, the size of the minority class is expanded by bootstrapping, which involves replacing the original minority class's data with new minority observations, or artificially creating new minority observations (Jain et al., 2021). Hybrid sampling combines oversampling and undersampling from the original sample and the rebalanced sample is approximately equal in size to the original sample.

There are two popular and well-known oversampling methods namely ROSE (Random oversampling examples), and SMOTE (Synthetic minority oversampling technique) in literature. SMOTE (Chawla, 2002) is an oversampling approach that generates new minority class instances at random from the sample's nearest minority class neighbors. This approach is done using the Euclidean distance between data points in feature space to determine nearest neighbors. Another important approach which is to generate artificial data based on sampling is the ROSE (Menardi and Torelli, 2014). This approach is aimed at oversampling the rare class by creating synthetic data points that are as similar as possible to the real ones with respect to a probability distribution centered on the selected sample (Liu, 2022). The imbalance ratio (*IR*) is a simple way to measure the unequal distribution of observations across classes. When the IR value is equal the 1, the dataset is perfectly balanced. The larger IR values indicate a larger difference in the class sizes. According to Chawla et al. (2002) the test set for the machine learning experiments must not include any “synthetic” samples. In this study, the dataset is split into a 70% training set and a 30% test set with a simple random sampling (SRS) strategy. It should be noted that oversampling approaches were only applied to the training set based on the above assumption.

3. Results and Discussions

The study aims to demonstrate the utilization of three ML methods for the analysis of both imbalanced and balanced data using several oversampling strategies. Therefore, prediction performances of ML algorithms namely NB, RF, and SVM are tested on a soil liquefaction dataset. In order to compare their performances, training sample sizes generated from ROSE, SMOTE, and simple oversampling (OVER) are also considered. Finally, for the assessment of the performances, Accuracy (*Acc*) and *Kappa* scores are utilized.

Firstly, liquefaction potential dataset is split into a 70% training set and a 30% test set with an SRS strategy. Then, oversampling methods were only applied to the training set. Training set obtained by SRS, balanced dataset obtained by oversampling applications, and *IR* values of training sets were given in Table 2. Oversampling methods namely SMOTE, ROSE, and OVER were used to increase the number of cases in a balanced way. As shown in the table, the training set (i.e., SRS) consisted of 158 liquefaction potential events which is 93 with value “Yes” and 65 with value “No”. The *IR* value after the sampling was found to be 1.43 and the result of this calculation indicates that the ratio of the training set sample size is imbalanced. On the other hand, the *IR* values are equal to 1 after applied oversampling strategies. Therefore, it can be clearly said that oversampling strategies prevent the data imbalanced. While Table 2 presents a brief view of the distributions of training data sampling ratios, Fig. 2 shows the change of “Yes” and “No” samples of the dataset after applying the oversampling methods in order to give a visual idea of the effect of these methods.

For a better understanding of the impact of the resampling method in optimal training sample size selection, the training sets obtained from each method were utilized to predict the test dataset using the NB classifier. After that, the performances of the models were compared by *Acc* and *Kappa* scores (Fig. 3).

Table 2. The ratio of the dataset with IR values

	Training Data Sampling Ratio		
Sampling Method	**"No" 0**	**"Yes" 1**	***IR***
SRS	65	93	1.43
Over	93	93	1
SMOTE	186	186	1
ROSE	78	80	1.03

The result showed that after oversampling strategies, each ones have a significant impact on the prediction performance. When the performance results were analyzed, the results of the prediction model obtained by SMOTE was found above 90% for *Acc* and above 80% for *Kappa*. On the other hand, the lowest

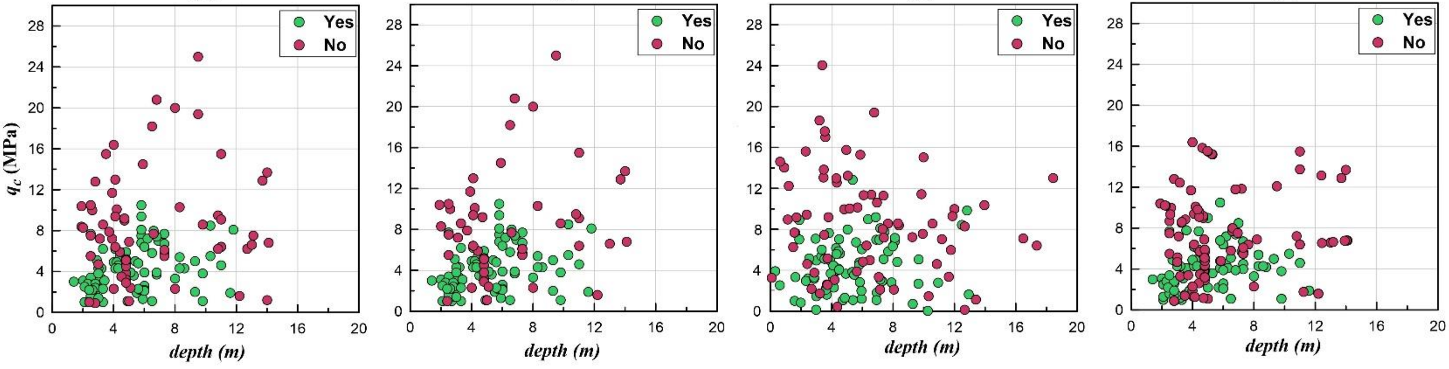

Figure 2. Variation of “Yes” and “No” samples in the training set with respect to different sampling methods (a) SRS, (b) OVER, (c) ROSE, and (d) SMOTE

performance scores (*Acc*= 87% and *Kappa*=72%) were obtained by traditional SRS strategy. Based on the performance results, it was decided to use the only SMOTE training set for the next applications.

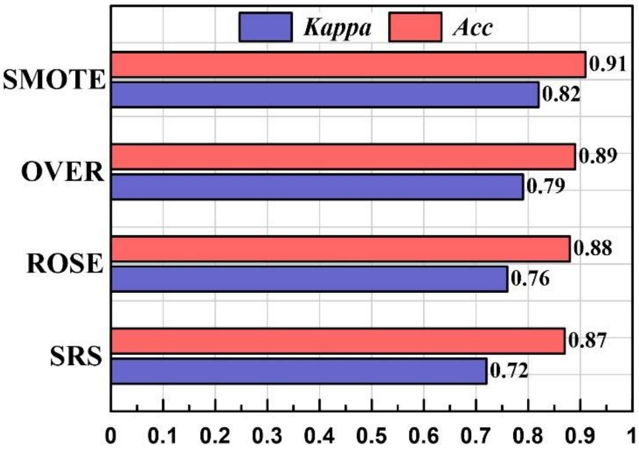

Figure 3. Comparison of performances of sampling methods

NB, RF, and SVM methods were trained by the training dataset obtained by SMOTE, and the model's performance results were obtained using the test dataset. In this purpose, accuracy performance metrics (i.e., *Acc*, *Kappa*, *Precision*, *Recall*, and *F-Measure*) obtained by confusion matrix (CM) were used to evaluate models.

To validate the predictive ability of the model with hyperparameter optimization, grid search (GS) with the k-fold Cross-validation (CV) technique was used. When making predictions on data that was not used during training, the k-fold CV procedure is used to estimate the performance of machine learning models. GS conducts an exhaustive search for the combination of parameters that maximizes the CV performance, according to the defined score function. In general, the choice of k is usually 5 or 10, but there is still no universal guideline or agreement for selecting the number of folds (k). Thus, 10-fold (i.e., k=10) CV was used in this study, and the optimum hyperparameter sets of models are given in Table 3.

The performance metric results were given with CM in Table 4. When the metric results according to *Acc* value between all models were analyzed, it was shown that SVM, RF, and NB methods were calculated as 0.9412, 0.9265, and 0.9118, respectively. When the metric results for SVM was examined in detail, which outperformed other models, the model had *Acc*, *Kappa*, and *F-Measure* values of 0.9412, 0.8799, and 0.9487, respectively. According to accuracy results, the SVM model was shown about 1.5% better results in the RF model and about 3% better results in the NB model. Furthermore, the model obtained with the SVM model trained by SMOTE training set was outperformed about 7% than the NB method using SRS training set. As a result, the performance of all models is quite acceptable when considering the sampling ratio strategies used. On the other hand, the performance metrics revealed that the SVM method trained by SMOTE sampling strategy showed better performance than the SVM trained by the conventional SRS method.

Table 3. The best set of hyperparameters

ML Method	Best Hyperparameters	Parameter Definition
SVM	'*C*': 1 '*gamma*': 0.3045 'kernel': radial	*C* (cost): Cost of constraints violation, *gamma*: regularization parameter
RF	'*ntree*': 500 '*mtry*': 4	*ntree:* number of trees *mtry:* number of features used to grow each tree
NB	*'usekernel': TRUE* *'adjust': 1* *'fL': 0*	*usekernel:* Allow using a kernel density estimate for continuous variables versus a Gaussian density estimate. *adjust:* adjust the bandwidth of the kernel density *fL:* Allowing to incorporate the Laplace smoother

4. Conclusions

In this study, different prediction models using SVM, RF, and NB were developed to predict the soil liquefaction. A total of 226 CPT data were used for the modeling of the prediction models then their performance results were compared each other using the five metrics. Moreover, three oversampling methods (OVER, SMOTE, and ROSE) were applied in this study to balance the training sets of the CPT dataset. The result showed that oversampling strategies have a significant impact on the prediction models, but the SMOTE one is highest than the others.

Table 4. Performance results of the SVM, RF, and NB models trained by SMOTE training set

SVM				RF				NB			
		Actual				Actual				Actual	
		No	Yes			No	Yes			No	Yes
Predicted	No	27	3	Predicted	No	26	3	Predicted	No	24	2
	Yes	1	37		Yes	2	37		Yes	4	38
Acc	:	0.9412		*Acc*	:	0.9265		*Acc*	:	0.9118	
Kappa	:	0.8799		*Kappa*	:	0.849		*Kappa*	:	0.8159	
Precision	:	0.9737		*Precision*	:	0.9487		*Precision*	:	0.9048	
Recall	:	0.9250		*Recall*	:	0.9250		*Recall*	:	0.9500	
F-Measure	:	0.9487		*F-Measure*	:	0.9367		*F-Measure*	:	0.9268	

Also, the result clearly showed that SVM with SMOTE model is a superior model than the rest with the highest accuracy. The order of the applied models' performance is SVM > RF> NB as per their Acc metrics over testing phase i.e., 94.12%, 92.65%, and 91.18%, respectively.

References

Adalier, K., & Elgamal, A. (2004). Mitigation of liquefaction and associated ground deformations by stone columns. *Engineering Geology*, 72(3-4), 275-291.

Allen, J. R. L. (1982). Sedimentary Structures: Their Character and Physical Basis. Volume II. Developments in Sedimentology, 30B, Amsterdam.

Amiri, M., Bakhshandeh Amnieh, H., Hasanipanah, M., & Mohammad Khanli, L. (2016). A new combination of artificial neural network and K-nearest neighbors models to predict blast-induced ground vibration and air-overpressure. *Engineering with Computers*, 32(4), 631-644.

Cetin, K. O., Seed, R. B., Der Kiureghian, A., Tokimatsu, K., Harder Jr, L. F., Kayen, R. E., & Moss, R. E. (2004). Standard penetration test-based probabilistic and deterministic assessment of seismic soil liquefaction potential. *Journal of Geotechnical and Geoenvironmental Engineering*, 130(12), 1314-1340.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321-357.

Chen, B., Xia, S., Chen, Z., Wang, B., & Wang, G. (2021). RSMOTE: A self-adaptive robust SMOTE for imbalanced problems with label noise. *Information Sciences*, 553, 397-428.

Demir, S., & Sahin, E. K. (2022). Comparison of tree-based machine learning algorithms for predicting liquefaction potential using canonical correlation forest, rotation forest, and random forest based on CPT data. *Soil Dynamics and Earthquake Engineering*, 154, 107130.

Douzas, G., & Bacao, F. (2017). Self-Organizing Map Oversampling (SOMO) for imbalanced data set learning. *Expert Systems with Applications*, 82, 40-52.

Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). Learning from imbalanced data sets (Vol. 10, pp. 978-3). Berlin: Springer.

He H., & Ma, Y. (2013) Imbalanced learning: foundations, algorithms, and applications. John Wiley & Sons, Inc., Hoboken, New Jersey.

He, S., Wu, J., Wang, D., & He, X. (2022). Predictive modeling of groundwater nitrate pollution and evaluating its main impact factors using random forest. *Chemosphere*, 290, 133388.

Hu, J., Zou, W., Wang, J., & Pang, L. (2021). Minimum training sample size requirements for achieving high prediction accuracy with the BN model: A case study regarding seismic liquefaction. *Expert Systems with Applications*, 185, 115702.

Jain, D., Mishra, A. K., & Das, S. K. (2021). Machine learning based automatic prediction of Parkinson's disease using speech features. *In Proceedings of International Conference on Artificial Intelligence and Applications* (pp. 351-362). Springer, Singapore.

Juang, C. H., Yuan, H., Lee, D. H., & Lin, P. S. (2003). Simplified cone penetration test-based method for evaluating liquefaction resistance of soils. *Journal of Geotechnical and Geoenvironmental Engineering*, 129(1), 66-80.

Kayen, R., Moss, R. E. S., Thompson, E. M., Seed, R. B., Cetin, K. O., Kiureghian, A. D., ... & Tokimatsu, K. (2013). Shear-wave velocity–based probabilistic and deterministic assessment of seismic soil liquefaction potential. *Journal of Geotechnical and Geoenvironmental Engineering*, 139(3), 407-419.

Koopialipoor, M., Fahimifar, A., Ghaleini, E. N., Momenzadeh, M., & Armaghani, D. J. (2020). Development of a new hybrid ANN for solving a geotechnical problem related to tunnel boring machine performance. *Engineering with Computers*, 36(1), 345-357.

Liu, J. (2022). Importance-SMOTE: a synthetic minority oversampling method for noisy imbalanced data. *Soft Computing*, 26, 1141–11631.

Menardi, G., & Torelli, N. (2014). Training and assessing classification rules with imbalanced data. *Data mining and knowledge discovery*, 28(1), 92-122.

Robertson, P. K., & Wride, C. E. (1998). Evaluating cyclic liquefaction potential using the cone penetration test. *Canadian Geotechnical Journal*, 35(3), 442-459.

Samui, P. (2008). Support vector machine applied to settlement of shallow foundations on cohesionless soils. *Computers and Geotechnics*, 35(3), 419-427.

Seed, H. B., & Idriss, I. M. (1971). Simplified procedure for evaluating soil liquefaction potential. *Journal of the Soil Mechanics and Foundations Division*, 97(9), 1249-1273.

Vluymans, Sarah. Dealing with Imbalanced and Weakly Labelled Data in Machine Learning Using Fuzzy and Rough Set Methods. Ghent University. Faculty of Medicine and Health Sciences; University of Granada. Department of Computer Science and Artificial Intelligence, 2018.

Wang, L., Wu, C., Tang, L., Zhang, W., Lacasse, S., Liu, H., & Gao, L. (2020). Efficient reliability analysis of earth dam slope stability using extreme gradient boosting method. *Acta Geotechnica*, 15(11), 3135-3150.

Wu, C., Fang, C., Wu, X., & Zhu, G. (2020). Health-risk assessment of arsenic and groundwater quality classification using random Forest in the Yanchi region of Northwest China. *Exposure and Health*, 12(4), 761-774.

Xie, Y., Ebad Sichani, M., Padgett, J. E., & DesRoches, R. (2020). The promise of implementing machine learning in earthquake engineering: A state-of-the-art review. *Earthquake Spectra*, 36(4), 1769-1801.

Zhang, W., Li, H., Li, Y., Liu, H., Chen, Y., & Ding, X. (2021a). Application of deep learning algorithms in geotechnical engineering: a short critical review. *Artificial Intelligence Review*, 54(8), 5633-5673.

Zhang, Y., Xie, Y., Zhang, Y., Qiu, J., & Wu, S. (2021b). The adoption of deep neural network (DNN) to the prediction of soil liquefaction based on shear wave velocity. *Bulletin of Engineering Geology and the Environment*, 80(6), 5053-5060.

Zhao, Z., Duan, W., & Cai, G. (2021). A novel PSO-KELM based soil liquefaction potential evaluation system using CPT and Vs measurements. *Soil Dynamics and Earthquake Engineering*, 150, 106930.

Zhou, J., Huang, S., Wang, M., & Qiu, Y. (2021). Performance evaluation of hybrid GA–SVM and GWO–SVM models to predict earthquake-induced liquefaction potential of soil: a multi-dataset investigation. *Engineering with Computers*, https://doi.org/10.1007/s00366-021-01418-3.