Combining SMOTE sampling and Machine Learning for Forecasting Wheat Yields in France

Amine Chemchem, Francois Alin, Michaël Krajecki

Combining SMOTE sampling and Machine Learning for Forecasting Wheat Yields in France

1st Amine Chemchem
University Of Reims Champagne-Ardenne
CReSTIC Lab
Reims, France
mohamed-lamine.chemchem@univ-reims.fr

2nd François Alin
University Of Reims Champ-Ardenne
CReSTIC Lab
Reims, France
francois.alin@univ-reims.fr

3rd Michaël Krajecki
University Of Reims Champ-Ardenne
CReSTIC Lab
Reims, France
michael.krajecki@univ-reims.fr

***Abstract*—This paper describes a method of predicting wheat yields based on machine learning, which accurately determines the value of wheat yield losses in France. Obtaining reliable value from yield losses is difficult because we are tackling a highly unbalanced classification problem. As part of this study, we propose applying the Synthetic Minor Oversampling technique (SMOTE) as a pretreatment step before applying machine learning methods. The approach proposed here improves the accuracy of learning and allows better results on the set of tests by measuring the operating characteristic of the ROC receiver. The comparative study shows that the best result obtained is 90.07% on the set of tests, obtained by hybridizing the SMOTE algorithm with the Random Forest algorithm. The results obtained in this study for wheat yield can be extended to many other crops such as maize, barley, ...**

***Index Terms*—Machine Learning , Knowledge Discovery , Smart Agriculture , Supervised Classification , Imbalanced Learning , Sampling Methods.**

I. Introduction

Wheat yield forecasts are a valuable source of information as decision-makers or commodity traders have the responsibility to adjust their import / export plans and prices. This information is also very useful for farmers to plan their wheat harvest and storage. In addition, yield forecasts can be integrated with crop insurance systems to cover the risk of significant losses due to adverse weather conditions.

Yield is the amount of grain harvested, usually expressed in tonnes per hectare. It depends on the characteristics of the region of culture (climate, geography, ...).

The aim of this study is to develop a machine learning model that can predict wheat yields in France as accurately as possible. The training data has been collected for 58 years and is published by Cland[1] as part of a challenge.

Given the highly unbalanced nature of this dataset, we have been led to adopt resampling methods as a pre-processing step to obtain a well-balanced and easier-to-process data set.

The main contributions of our paper are : (1) to the best of our knowledge, the first hybridization of SMOTE and machine learning methods for forecasting wheat yield on a real world dataset; and (2) The comparative study of machine Learning approaches, tuned and trained on the nvidia DGX1 device, which can train model 58 times faster as compared to a classical CPU [1].

The remainder of the paper is organized as follows. Section II reviews some researches in artificial intelligence and machine learning applied on smart agriculture. Section III describes the data and the experimental setup of the experiments presented in Section IV. Finally, section V concludes the paper with some perspectives.

II. Literature Survey

In the field of artificial intelligence for agriculture, the authors of [2] combine decision trees based on the ID3 algorithm and farmers' knowledge to develop an expert system. This expert system aims to provide advice on tomato harvesting. The user has an online interaction with the expert system; he must answer the questions asked by the expert system. Depending on the response of the user, the expert system detects the disease and displays its control of the disease. This expert system on tomato crops deals with different varieties of tomato crops. The identification of various diseases usually occurs on tomato crops depending on the symptoms. This rule-based expert system validates the symptoms of tomato harvesting using ID3 algorithm techniques and some optimization algorithms.

Several applications of Data Mining techniques can be found in the field of agriculture. Researchers at [3] have implemented the K-Means algorithm to predict pollution in the atmosphere. The K Nearest Neighbor method is applied [4] to simulate daily precipitation as well as other meteorological variables. In [5], different possible weather scenarii changes are analyzed using support vector machines. The grouping techniques are applied in the classification of apples before their marketing [6]. These techniques are also for the detection of weeds in precision farming [7].

In Machine Learning and Deep Learning, some authors [8] introduce a precise and inexpensive method for predicting crop yields using public remote sensing data. The proposed approach is based on a new dimensionality reduction technique that makes it possible to define a convolutional neural network or a long-term memory network and to automatically learn useful functionalities even when the labeled training data is scarce. In addition, the authors incorporate a component of

[1]https://cland.lsce.ipsl.fr/index.php/workshops/forecasting-crop-yields/28-yield-forecasting

TABLE I
DATA DESCRIPTION

Attribute	Description
Class	Value equal to 1 in case of severe loss of wheat yield and zero otherwise. This is the target variable to predict.
Year harvest	year (anonymous) harvest (1 to 58).
NUMD	number indicating the France department (from 1 to 94).
ETP 1 ... ETP 12	Potential mean monthly evapotranspiration by year and department (1 = January, 12 = December).
PR 1 ... PR 12	Monthly accumulated precipitation per year by department.
RV 1 ... RV 12	Average monthly radiation per year and per department.
SeqPR1 ... SeqPR12	Number of rainy days per year per department.
Tn 1 ... Tn 12	Minimum monthly average daily temperature by year and by department.
Tx 1 ... Tx 12	Maximum monthly average daily temperature by year and by department.
Tn17.1 1 ... Tn17.1 12	Number of days where the minimum daily temperature is below -17 degrees C for each month per year and per department.
Tx010 1 ... Tx010 12	Number of days where the maximum daily temperature is between zero and 10 degrees C for each month per year and per department.
Tx34 1 ... Tx34 12	Number of days where the daily maximum temperature is above 34 degrees C for each month per year and per department.

the Gaussian process to explicitly model the spatio-temporal structure of data and improve accuracy. The evaluation of this approach to county-level soybean yield forecasting in the United States shows that it outperforms competing techniques. In [9], the authors review several strategic applications of machine learning in maize breeding. Quantitative mapping of traits and selection of populations at the genome level are some of the key areas currently addressed in the literature. The results show that machine learning algorithms are a serious alternative to traditional statistical techniques applied to maize, as well as linear mixed models introduced more recently.

III. RESEARCH MATERIALS & METHODOLOGY

A. Research Materials

The data used in this study resumes wheat yield in France, which were been compiled over a continuous 58 years period. The dataset was published by Cland institute[2] as part of challenge called "Crop Data Challenge 2018". This benchmark is taken in ninety three input variables which are described in table I

In the same way of our previous studies in [10] [11], the presented machine learning approaches are implemented, optimized and evaluated on the Nvidia DGX1[3] server: a system specially designed for deep learning and artificial intelligence, which includes eight Tesla V100 GPUs, connected to each other through an NVlink network and supporting up to 40 GB/s bidirectional bandwidth.

B. Research Methodology

We begin by dividing the dataset into a learning subset and a test subset (80% and 20%, respectively). Then, we perform resampling using techniques such as subsampling or the Synthetic Minor Oversampling technique (SMOTE) to obtain a perfectly balanced training set. Finally, we implement various popular machine learning approaches to compare their performance. It should be noted here that, for each learning approach, the best hyper parameters are chosen using cross-validation grid search [12]. The best model is chosen on the basis of precision criteria to predict new data. All these steps are shown in Figure2.

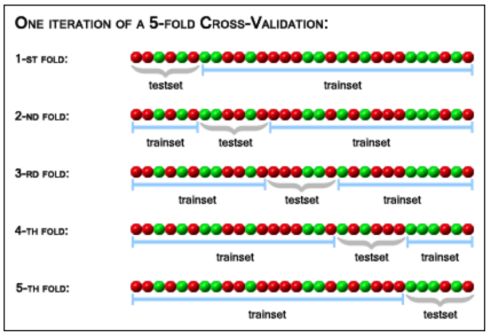

Fig. 1. k-fold cross validation example

1) Splitting the dataset into Train/Test with cross-validation : Basically, the first step in machine learning is to divide the dataset into two sets one called training set and another test set. The first contains the data with well labeled examples, it is used to build our model, when the second one is used to test the performance of the model.

Generally, we split our dataset according to 80/20 rule i.e 80% of dataset goes to training set and 20% goes to test set. Random Train/Test split method provides high variance estimate, since changing which observations or examples happens to be in testing dataset can significantly change testing accuracy.

To avoid this drawback, the cross validation method is used. It consists on splitting the dataset into bunch of train/test splits, calculate their training accuracy and average the results together.

The k-fold cross validation is the common type of cross

[2]Cland institute : funded by the French government in 2017 to perform the research urgently needed on land-management solutions for managing the ecological and energy transitions of the 21st century.

[3]https://labs.ovh.com/nvidia-dgx-1

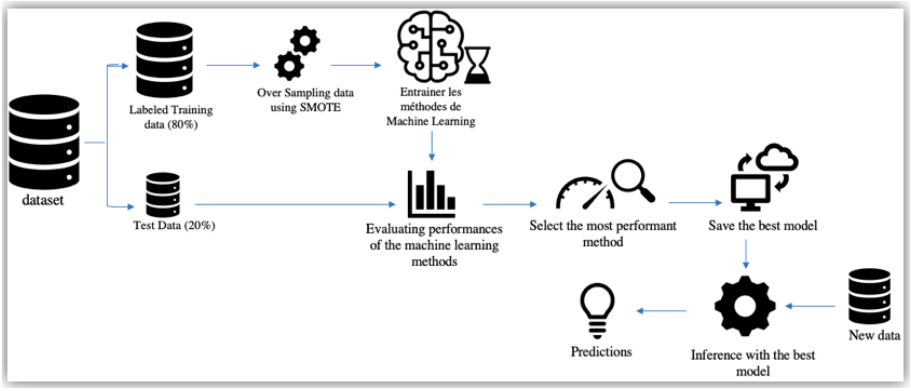

Fig. 2. General Workflow

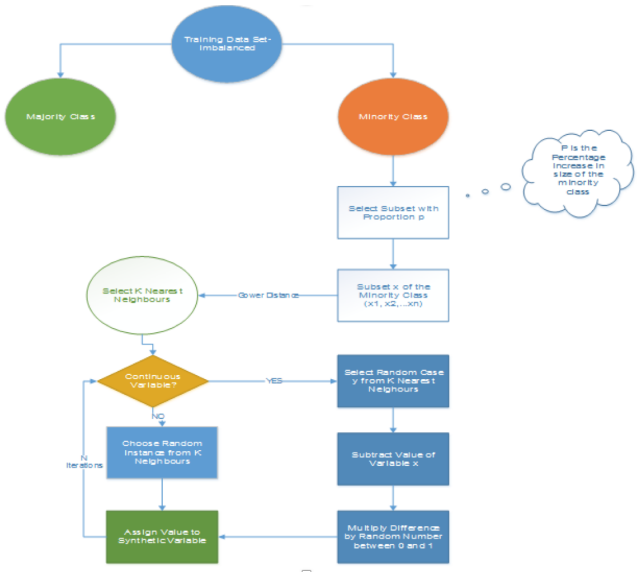

Fig. 3. Synthetic Minority Oversampling Organigram [18]

validation, it is used to avoid the over fitting in a predictive model, particularly in a case where the amount of data may be limited. We start by fixing the parameter 'K', which represents the number of folds (or partitions) of the data, then we run the analysis on each fold, after that, we average the overall error estimate. Figure 1 shows an example of k-fold cross validation with k equals to 5.

2) Over-sampling the dataset using SMOTE: The dataset is strongly unbalanced, as shown in figure 6. We notice here that, the class value is equal to 1 in case of severe loss of wheat yield, and it is equal to 0 otherwise. This represents a kind of comparison between the statistical yield prediction model and the real yield.

Generally, machine learning algorithms have trouble learning when one class dominates the other. For this reason, we apply Synthetic Minority Over-sampling Technique (SMOTE) [14].

In SMOTE, the minority class is over-sampled by introducing synthetic instances where each minority class sample is taken. The generated data are inserted along the line segments joining some of the k-nearest neighbors of the minority class. Neighbors are randomly chosen from k-nearest neighbors depending upon the amount of over-sampling that is required. Five nearest neighbors are currently used in the implementation of SMOTE [15] [16].

In short SMOTE algorithm can be stated in the steps as, taking the difference between the feature vector (minority class example) under consideration and its nearest neighbor (minority class examples) and then multiplying this difference by a random number between 0 and 1. Furthermore, adding the difference calculated in previous step to the feature vector as a result creating a new feature vector [17], the full algorithm is represented in figure 3.

For instance, if we consider a sample (6,4) for which k-nearest neighbors are being identified. Let (4,3) is one of its k-nearest neighbors.

Let: $f1_1 = 6$, $f2_1 = 4$, $f2_1$ - $f1_1$ = -2.

$f1_2 = 4$, $f2_2 = 3$,$f2_2$ - $f1_2$ = -1.

The new samples will be generated as : (f1',f2') = (6,4) + rand(0-1) * (-2,-1).

Rand(0-1) generates a random number between 0 and 1.

IV. Results and Analysis

In figures 4, and 5 we visualize some data for a best analyse and comprehension. For instance figure 4 shows how predictions of yield loss change per year.

We note that the in majority of cases, losses of yield are below of 45%, which indicates that the yield prediction model is not very wrong. Moreover, there is no apparent relationship between a year and the following year in terms of loss prediction, since there is no stability in the graph.

In the same way, if we want to show the yield prediction losses by department, we get figure 5. We notice that there are no data collected from the department 45.

A. Machine Learning Approaches Comparison

We compare here most of machines learning approaches like: K-Nearest Neighbors Classifier, Support Vector machine, Gradient Boosting Classifier, Ada Boost Classifier, Decision Tree Classifier, Random Forest Classifier, Multi Layer Perceptron Classifier, Naive Bayes Classifier, and XGBoost Classifier.

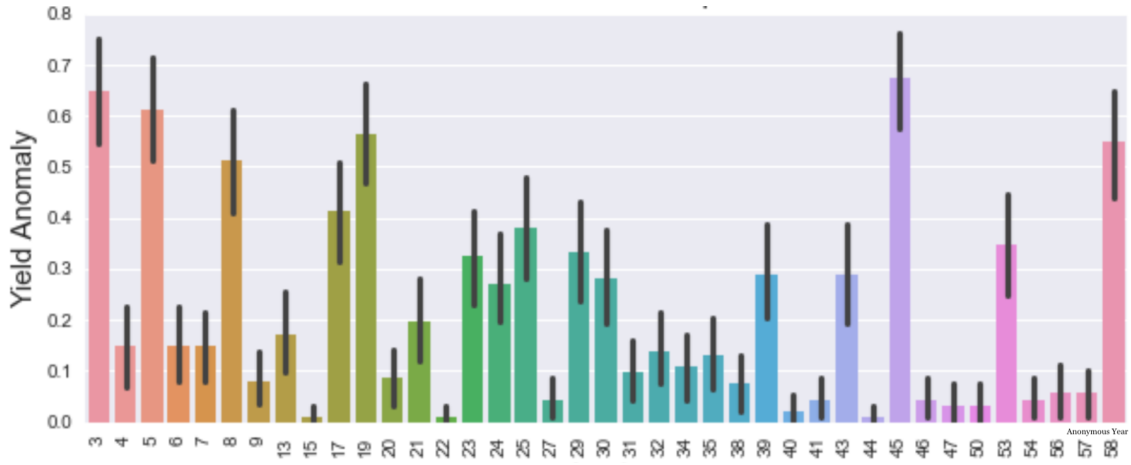

Fig. 4. Losses of harvest yield per anonymous year

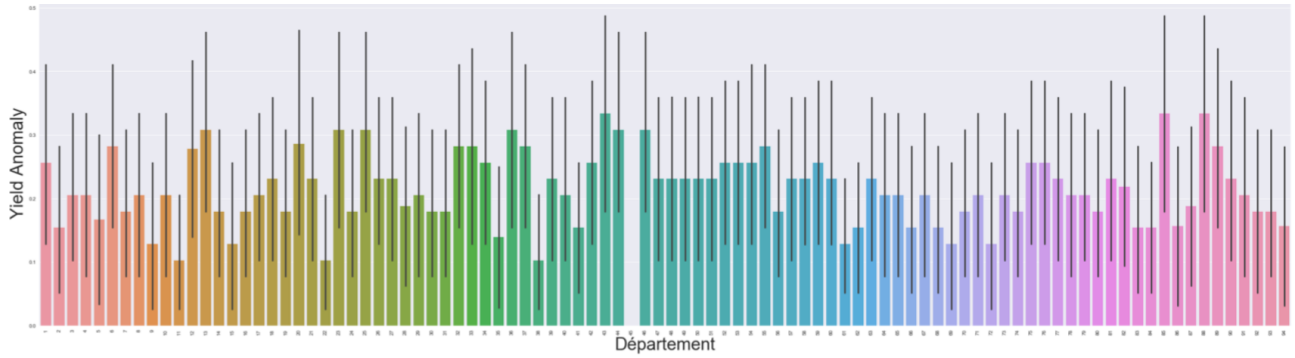

Fig. 5. Losses of harvest yield per France department

For the performance evaluation, we use AUC (Area Under The Curve) ROC (Receiver Operating Characteristics) curve. It is one of the most important evaluation metrics for checking any classification models performance. It is also written as AUROC (Area Under the Receiver Operating Characteristics) [19]. We calculate also the accuracy and the F-score of each approach to validate the obtained results with the AUCROC measure. The F-score also called F-measure is based on the two primary metrics : precision and recall. Given a subject and a gold standard, precision is the proportion of cases that the subject classified as positive that were positive in the gold standard. It is equivalent to positive predictive value. Recall is the proportion of positive cases in the gold standard that were classified as positive by the subject. It is equivalent to sensitivity. The two metrics are often combined as their harmonic mean [20], the formula can be formulated as follows:

$$F = \frac{(1+\beta^2) \times recall \times precision}{(\beta^2 \times precision) + recall}$$

$$Precision = \frac{TP}{TP+FP}, Recall = \frac{TP}{TP+FN}$$

Where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives

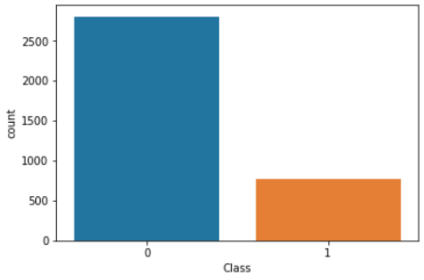

Fig. 6. Training dataset distribution

and FN is the number of false negatives. The F-measure can be used to balance the contribution of false negatives by weighting recall through a parameter $\beta \geq 0$ [21]. In our case β is set to 1, F1-score is than equal to:

$$F1\_score = \frac{2 \times recall \times precision}{precision + recall}$$

As reported in figure 6, the dataset is strongly unbalanced. For this reason we propose to perform multiple strategies of comparison : the first one without making any preprocessing on the dataset. The second way is reached by the application of an Under-Sampling approach, which consists on dropping data

TABLE II
F-SCORE & PRECISION COMPARISON OF THE MACHINE LEARNING APPROACHES

ML Approach	Without Sampling		Under Sampling		SMOTE Sampling	
	Accuracy	F1-Score	Accuracy	F1-Score	Accuracy	F1-Score
K-Nearest Neighbors	84.30%	0.84	75.42%	0.75	76.48%	0.78
Support Vector Machine	81.50%	0.73	56.21%	0.41	79.61%	0.71
Gradient Boosting	84.51%	0.82	77.11%	0.77	82.30%	0.83
Ada Boosting	82.58%	0.81	73.16%	0.73	77.49%	0.79%
Decision Tree	79.35%	0.80	69.49%	0.69	79.39%	0.80%
Random Forest	83.65%	0.81	74.94%	0.75	84.09%	**0.84**
Multi Layer Perceptron	81.07%	0.79	53.10%	0.49	76.37%	0.78
Naive Bayes	81.50%	0.74	56.49%	0.51	58.11%	0.62%

TABLE III
PERFORMANCES COMPARISON OF THE MACHINE LEARNING APPROACHES

ML Approach	Without Sampling		Under Sampling		SMOTE Sampling	
	Training Score	AUC ROC	Training Score	AUC ROC	Training Score	AUC ROC
K-Nearest Neighbors	89.18%	0.67	84.71%	0.74	90.96%	0.84
Support Vector Machine	100%	0.50	100%	0.50	100%	0.50
Gradient Boosting	91.03 %	0.68	93.27%	0.75	91.36%	0.84
Ada Boosting	85.22%	0.67	82.30%	0.72	81.20%	0.73
Decision Tree	100%	0.71	100%	0.66	100%	0.69
Random Forest	99.40%	0.72	99.78%	0.73	99.92%	**0.90**
Multi Layer Perceptron	97.39%	0.52	69.70%	0.65	72.37%	0.71
Gaussian Naive Bayes	76.54%	0.61	70.70%	0.68	70.82%	0.70
Bernoulli Naive Bayes	78.01%	0.50	61.50%	0.52	60.32%	0.55

rows randomly of the majority class. The third comparison is realised by the application of SMOTE approach, which generates new synthetics data as explained in the subsection III-B.

The table III summarizes the results obtained by the different machine learning approaches, and by applying the three sampling methods. For each algorithm, in addition to the AUC ROC scores, its training score is reported.
We note here that the results mentioned are the best scores of each method after boosting its parameters. The process of parameter tuning is achieved using the toolbox "sklearn.grid_search.GridSearchCV", it consists on selecting the values for a models parameters that maximize the accuracy of the model [22].

Figure 7 shows a visualization of the comparison study. We remark that the best results are given by applying the SMOTE sampling method. We notice that the SVM algorithm obtains 100% in training score and 50% in test score, and this in all sampling cases. This is a case of over fitting, probably due to the fact that the learning base contains a lot of features variables and not too many examples of data rows.

We remark also that, the best result on AUC ROC score is obtained by the Random Forest Approach when Sampling with SMOTE Algorithm. In addition, this approach achieve the best accuracy and the best F-score on the test set like shown in tableII. We can explain this result by the fact that, in its principle, Random Forest select some features to construct the decision trees, unlike SVM which deals with all features. This selection allows to avoid a possible over fitting case, and also can help to get a best accuracy when choosing the most important features.The results show that Random Forest reaches 90,07% on AUC ROC score, while the best of the other approaches gets 84,69%.

V. CONCLUSION

In this paper we compared different supervised learning methods such as support vector machine, multilayer perceptron, K-Nearest Neighbors, gradient boosting, Ada Boost, decision tree , random forest and Naive Bayes method to predict wheat yield from meteorological data. We have shown that it is the Random Forest method that has the highest AUC ROC score in front of all the others with a score of 90.07% positive prediction. The proposed method is very useful for farmers because knowledge of pre-harvest performance allows them to have better crop planning, better inventory management and optimize their contracts (purchase and sale of grain). Performance forecasts are also strategic information used by international actors. Abundant or, in contrast, low crop forecasts can have a significant impact on world agricultural market prices. Nevertheless, through this study we can affirm that departments that have had the same weather conditions will have similar yields. For the future, we plan to cross this database with satellite data for better accuracy. We will also try to predict the model two or three months before the harvest date. Of course, this approach could apply to various crops such as barley, corn, etc.

REFERENCES

[1] Shafique, Muhammad, et al. "Adaptive and energy-efficient architectures for machine learning: Challenges, opportunities, and research roadmap." VLSI (ISVLSI), 2017 IEEE Computer Society Annual Symposium on. IEEE, 2017.

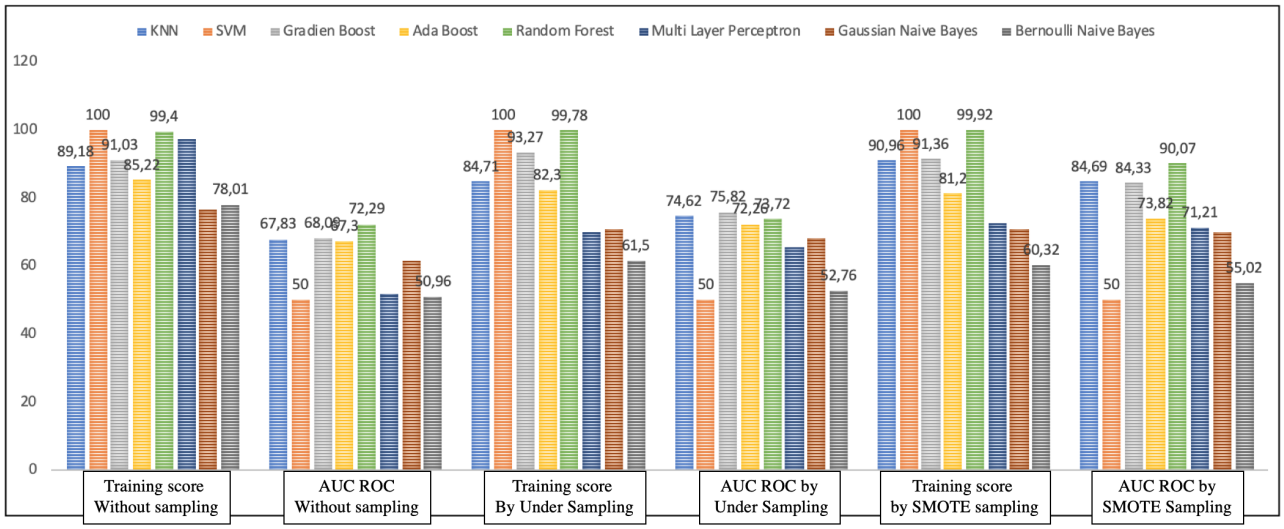

Fig. 7. Performance Comparison of Machine Learning Approaches

[2] Babu MSP, Ramana Murty NV, Narayana SVNL. A web based tomato crop expert information system based on artificial intelligence and machine learning algorithms. International Journal of Computer Science and Information Technologies. 2010; 1(3):15

[3] Jorquera H, Perez R, Cipriano A, Acuna G, "Short Term Forecasting of Air Pollution Episodes", In: Zannetti P (eds) Environmental modeling , WIT Press, UK, 2001.

[4] Rajagopalan B, Lall U, "A K-Nearest Neighbor Simulator for Daily Precipitation and Other Weather Variables", Wat Res Res 35(10), 1999, pages : 3089-3101.

[5] Tripathi S, Srinivas V V, Nanjundiah R S, "Downscaling of Precipitation for Climate Change Scenarios: A Support Vector Machine Approach", J Hydrol, 2006, pages : 621-640.

[6] Leemans V, M F Destain, "A Real Time Grading Method of Apples Based on Features Extracted from Defects", J. Jood Eng., 2004, pages : 83-89.

[7] Tellaeche A, X P Burgos Artizzu, G Pajares, A Ribeiro, "A Vision-Based Classifier for Weeds Detection in Precision Agriculture through the Bayesian and Fuzzy K-Means Paradigms", Adv.Soft. Comp., 2008, pages : 72-79.

[8] You, J., Li, X., Low, M., Lobell, D., & Ermon, S. (2017, February). Deep Gaussian Process for Crop Yield Prediction Based on Remote Sensing Data. In AAAI (pp. 4559-4566).

[9] Ornella L, Cervigni G, Tapia E. Applications of Machine Learning for Maize Breeding. In: Venkateswarlu B, Shanker AK, Shanker C. Book chapter of Crop stress and its man- agement: Perspectives and Strategies, Springer, New York, USA. 2012; 129.

[10] Chemchem, Amine, François Alin, and Michael Krajecki. "Improving the Cognitive Agent Intelligence by Deep Knowledge Classification." International Journal of Computational Intelligence and Applications (2019): 1950005.

[11] Chemchem, Amine, François Alin, and Michael Krajecki. "Deep Learning and Data Mining Classification through the Intelligent Agent Reasoning." 2018 6th International Conference on Future Internet of Things and Cloud Workshops (FiCloudW). IEEE, 2018.

[12] Kohavi, Ron. "A study of cross-validation and bootstrap for accuracy estimation and model selection." Ijcai. Vol. 14. No. 2. 1995.

[13] Guo, William W., and Heru Xue. "Crop yield forecasting using artificial neural networks: A comparison between spatial and temporal models." Mathematical Problems in Engineering (2014).

[14] Chawla, Nitesh V., et al. "SMOTE: synthetic minority over-sampling technique." Journal of artificial intelligence research 16 (2002): 321-357.

[15] Chen, Y. "Learning classifiers from imbalanced, only positive and unlabeled data sets". Project Report for UC San Diego Data Mining Contest. Computer Science Department, Iowa State University, Ames, IA. (2008).

[16] He, Haibo, and Edwardo A. Garcia. "Learning from imbalanced data." IEEE Transactions on Knowledge & Data Engineering 9 (2008): 1263-1284.

[17] Jishan, Syed Tanveer, et al. "Improving accuracy of students final grade prediction model using optimal equal width binning and synthetic minority over-sampling technique." Decision Analytics 2.1 (2015).

[18] Upasana, "How to handle Imbalanced Classification Problems in machine learning?". Analytics Vidhya. https://www.analyticsvidhya.com/blog/2017/03/imbalanced-classification-problem/. (2017).

[19] Huang, Jin, and Charles X. Ling. "Using AUC and accuracy in evaluating learning algorithms." IEEE Transactions on knowledge and Data Engineering 17.3 (2005): 299-310.

[20] Hripcsak, George, and Adam S. Rothschild. "Agreement, the f-measure, and reliability in information retrieval." Journal of the American Medical Informatics Association 12.3 (2005): 296-298.

[21] Chemchem, Amine, and Habiba Drias. "From data mining to knowledge mining: Application to intelligent agents." Expert Systems with Applications 42.3 (2015): 1436-1445.

[22] Ndiaye, Eugene, et al. "Safe Grid Search with Optimal Complexity." arXiv preprint arXiv:1810.05471 (2018).