IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Revue d'Intelligence Artificielle
Vol. 36, No. 6, December, 2022, pp. 855-862
Journal homepage: http://iieta.org/journals/ria

Impact of Recursive Feature Elimination with Cross-validation in Modeling the Spatial Distribution of Three Mosquito Species in Morocco

Meriem Douider[1*], Ibrahim Amrani[2], Thomas Balenghien[3,4,5], Amal Bennouna[6], Mounia Abik[1]

[1] Advanced Digital Enterprise Modeling and Information Retrieval Laboratory, ENSIAS, Mohammed V University, Rabat 10100, Morocco
[2] Smart Systems Laboratory, ENSIAS, Mohammed V University, Rabat 10100, Morocco
[3] UMR ASTRE, CIRAD, Rabat 10101, Morocco
[4] UMR ASTRE, CIRAD, University of Montpellier, CIRAD, Montpellier, France
[5] Unit Parasitology and Parasitic Diseases, Institute of Agronomy and Veterinary II, Rabat 10100, Morocco
[6] Department of Virology, Pathogen Discovery Laboratory, Pasteur Institute, Paris 75015, France

Corresponding Author Email: meriem_douider@um5.ac.ma

https://doi.org/10.18280/ria.360605

Received: 8 November 2022
Accepted: 17 December 2022

Keywords:
feature selection, data preprocessing, modeling, improved performance, mosquito

ABSTRACT

Many studies in ecology are interested in characterizing the ecological factors; determining the distribution of animal species. The classical approach consists in identifying the combination of ecological factors that allow reproducing observations of the presence and absence of the species of interest. The major difficulty lies in the imbalance between a considerable quantity of ecological factors to be tested and a relatively limited number of presence/absence observations. Selection of the most influential ecological features is a classical data pre-processing strategy that aims to overcome this imbalance and improve model performance. In this paper, we applied recursive feature elimination with cross-validation (RFECV) approach on presence/absence mosquito data in Morocco; to select optimal subsets of ecological features, in order to improve the performance of the predictive models. This method demonstrated the best ability to improve the performance of the predictive models, and can be recommended as a modeling improvement technique for large datasets.

1. INTRODUCTION

Machine learning is a subdomain of research in artificial intelligence (AI) that is developing rapidly and on which most AI applications are based. It offers a wide variety of algorithms and tools that are capable of learning autonomously, improving the accuracy of models, and making predictions while acquiring knowledge from real data without human intervention [1]. Machine learning algorithms have become increasingly important tools for addressing a variety of questions related to ecology, biogeography, conservation biology, and the consequences of climate change. These tools are, for example, classically used in the predictive modeling of mosquito species distribution [2, 3].

Morocco, due to its geomorphology and climate, is a country with a high diversity of flora and fauna, which makes it a privileged region for mosquitoes. Modeling the distribution of mosquitoes in Morocco is important for several reasons [4, 5]: First, mosquitoes are important vectors of disease, and understanding their distribution, abundance, and behavior in different regions can help public health officials develop strategies to control and prevent the spread of mosquito-borne diseases. Second, the diversity of mosquito species in Morocco is relatively high, with 43 different species known in the country. Finally, climate change is affecting the distribution and abundance of mosquitoes in many parts of the world, and studying mosquitoes in Morocco can help researchers understand how these changes affect the distribution and behavior of mosquito populations in the region.

The traditional approach, which is used in species distribution modeling, is to identify relationships between the known occurrence of a species (presence/absence) and ecological data (can be used meteorological data, topographical data, or vegetation characteristics). Then use these relationships to make predictions for unsampled areas of the study region. This approach has been used in several research studies to model mosquito distribution in several countries:

In the Netherlands, two studies [2, 6] were realized in 2015: they concentrated on mosquito occurrence data and 24 environmental features, including temperature features, vegetation index, infra-red index, precipitation features, population density, land cover, and digital elevation model. The modeling results showed the efficiency of the random forest model compared to two other models based on the results of sensitivity and specificity of these models, and the ten-best features of the best models were identified.

In Senegal in 2018 [3], a study of the distribution of a group of mosquito species was performed on an abundance dataset. Three different modeling approaches were compared to analyze the relationship between species abundance and 22 environmental features. These features can be classified into five groups (temperature, vegetation index, precipitation, land cover, and livestock density). Based on the random forest models, which provided better estimates of abundance,

environmental and climatic features that influence species abundance were determined.

In Germany in 2018 [7], a new mosquito modeling approach was applied to a dataset of mosquito abundance with eight meteorological features. This approach consists of combining several learning algorithms to improve the modeling performance. Based on the evaluation results, a specific combination of models can predict the distribution of mosquito species more effectively than a single model or a random combination of models.

In Morocco in 2021 [5], mosquito distribution modeling was performed on only presence record data with 20 environmental characteristics, including elevation, temperature, and precipitation features. The maximum entropy model was used to generate predictive models of potential mosquito distribution. The results obtained in this work can contribute to a better understanding of the potential distribution of each species and strengthen monitoring efforts in areas identified as high risk.

In the previous studies, it has been seen that the authors have applied a variety of algorithms to model the mosquito data but with a limited number of environmental features, which limits the overall understanding of their distribution. Another limitation is that the work done in Morocco on mosquito modeling is restricted due to a lack of absence record data. The present study will be the initial modeling of a new dataset on three species of mosquitoes in Morocco; prepared by specialists in entomology. The dataset available in this study contains a rich set of environmental features (225 environmental characteristics) compared to previous work and has the advantage of having information on the presence and absence of mosquitoes. This database is more detailed, containing many new features compared to previous work databases in the temperature, precipitation, and vegetation groups. In addition, it contains new groups of features such as animal distribution, wind speed, and water vapor. This diversity of environmental features will allow us to improve the modeling results and to understand more clearly the factors that influence mosquito distribution.

The presence/absence mosquito datasets often contain redundant, irrelevant, and noisy data: a species may be both present and absent at two nearby stations with the same ecological characteristics, or a species may be absent at a station with favorable ecological characteristics. This characteristic of this type causes an over-fitting of the model and increases the error rate of the learning algorithm. Well, to manage these problems and effectively use machine learning algorithms, data preprocessing is a crucial step. Feature selection [8] is one of the most common and important techniques in data preprocessing: it aims to select the non-redundant most relevant features from the original set so that they can be used in the construction of new models. This technique accelerates the modeling algorithms, improves the accuracy of predictions, and facilitates interpretation [9]. In the domain of modeling mosquito distribution data, feature selection techniques also help in the identification of favorable or unfavorable ecological factors for the development of these insects and consequently in the ecological interpretation of modeling results.

In general, feature selection methods can be classified into filtering, wrapper, and embedded methods [10]. Filter methods select features from the dataset without the use of a machine learning algorithm. Wrapper methods use both a learning algorithm and an evaluation criterion; they select the combination of features that gives the optimal results for the learning algorithm. In embedded methods, the feature selection algorithm is related to the learning algorithm, which thus integrates its own feature selection methods.

The objective of this study was to compare the capacity of the RFECV selection technique in order to select the best subset of features relative to other selection techniques using different machine learning algorithms on mosquito distribution data. The models obtained using the selection techniques were evaluated by cross-validation using three criteria: accuracy, Matthews Correlation Coefficient (MCC), and area under the curve ROC (AUC).

The results of this study demonstrated the efficiency of the RFECV technique in the selection of the best subsets of features and the improvement of the modeling performance. Indeed, the models obtained using the RFECV selection outperformed the models applied on all features, as well as the models obtained using a group of selection techniques, which leads to recommend this technique to be used for the improvement of modeling and the selection of features in the domain of data processing.

The remainder of the paper is organized as follows. Section 2 presents an overview of the selection and modeling techniques used. Section 3 focuses on the methodology followed for the application of the different methods. Section 4 includes the comparison and discussion of the results. Conclusion and future work are shown in section 5.

2. FEATURE SELECTION TECHNIQUES AND LEARNING ALGORITHMS CONSIDERED IN THIS STUDY

2.1 Feature selection techniques

2.1.1 Recursive feature elimination with cross-validation (RFECV)

The RFECV is a wrapper feature selection algorithm. Its principle is to select the features that have the greatest impact on the modeling improvement. It uses an iterative process of descending elimination for feature selection. This elimination is based on the importance of features, which can be calculated using a learning model that provides information about the importance of each feature [11, 12].

Algorithme 1: RFECV

1. *Choose a learning algorithm*
2. *Train/test this algorithm using cross-validation with all p features of the dataset*
3. *Calculate the performance of this model*
4. *Sort the p features by levels of importance and eliminate the least important feature*
5. *For i descending from p-1 to 1do:*
6. *Train/test the model using cross-validation with the most important i features*
7. *Calculate the performance of this model*
8. *Sort the i features by levels of importance and eliminate the least important feature*
9. *End*
10. *Display the features chosen for the best-performing model*

2.1.2 Feature selection according to importance "M10"

This technique is based on the importance of features in the

learning model. A modeling, called complete modeling, is first performed using all features, with a selected set of learning models and cross-validation. Then, the importance of each feature for the best model is calculated. After sorting in descending order of importance, the first features are selected as the most relevant features. In this study, we chose to select 10 features; a threshold regularly used in different research works [2, 3, 6].

2.1.3 ReliefF

ReliefF [13] is an improved algorithm of the original algorithm Relief [14], which calculates an approximate statistic for each feature. It estimates the quality of a feature based on how well randomly chosen values distinguish the nearby instances of the same class and the other class [15].

2.1.4 Info gain

Information gain is a filtering technique based on entropy to measure the quantity of information for each feature [16]. In this approach, the features selected are those that obtain the highest values of the information gain.

2.1.5 Correlation

The correlation coefficient measures the relationship between the dependent variable (presence or absence of a mosquito species) and each explanatory variable (environmental data). It can be used to feature selection as a filtering approach [17].

The literature often uses the Pearson correlation coefficient for feature selection [18]. This coefficient ranges from -1 to 1 and appreciates the existence of a linear relationship between the two variables.

In this paper, we also used a coefficient based on the ratio of inertia which is derived from the decomposition equation of the total inertia. It ranges between 0 and 1 and appreciates the homogeneity of the partition of each explanatory variable by the presence-absence variable.

Formula:

$$\rho = \frac{Between\ inertia}{Total\ inertia} = \frac{\sum_{k=1}^{2} n_k(\bar{x}_k - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad (1)$$

where, x_i is the i-th observation of an explanatory variable with mean $\bar{x}$ and $\bar{x}_k$ is the mean of this explanatory variable in the class with effective n_k.

2.2 Learning algorithms

2.2.1 Logistic regression (LG)

Logistic Regression is a predictive model developed in 1944 by Joseph Berkson. It aims to find the relationship between a set of explanatory variables with numerical or categorical values and a binary or multinomial categorical target variable. It is simply a non-linear transformation of Linear Regression, where we try to predict a class instead of a continuous numerical value.

Logistic Regression is one of the most common multivariate analysis models used in several application domains (for example; in epidemiology: [19], and in studies of mosquito species distribution: [2, 7]).

2.2.2 Gaussian naïve bayes (GN)

Gaussian Naïve Bayes is a variant of Naïve Bayes. Naïve Bayes methods are a set of supervised Machine Learning classification algorithm; based on Bayes' theorem with strong assumptions of independence between features. The use of this algorithm to discriminate between two types of cancer has shown very satisfactory results [20].

2.2.3 K nearest neighbors (KNN)

It is a supervised learning method that we can use for classification and regression. To predict the class of a new case, the algorithm finds the majority class of the K nearest neighbors. The method uses two parameters: the number K and the similarity function that determines the neighbors of each new case [21].

2.2.4 Random forest (RF)

Random Forest is a supervised learning algorithm introduced by Breiman [22]. It is built from several basic models consisting of Decision Trees, which are merged to obtain a more accurate and stable prediction. Each tree is constructed from a randomly generated sample of the training set [23].

2.2.5 Gradient boosting & XGBoost

Gradient Boosting (GB) is a machine learning technique developed by Friedman [24]. This method consists of running a series of learning algorithms, where each model is built on the residuals of the previous model. The predictive model commonly used with Gradient Boosting is the decision tree [25].

XGBoost (XG) is a specific implementation of the Gradient Boosting method that uses more accurate approximations to identify the best tree model. It is a model widely used by data scientists to address many learning challenges [26, 27].

3. EXPERIMENTAL DESIGN

3.1 Data description

The presence/absence observation data include occurrence data of different mosquito species collected in 366 sites (entomological data) in Morocco and more than 225 ecological variables associated.

The entomological data derive from two sources: data from larval surveys conducted in 2015-2016 [28]; generating presence and absence data, and data from the Moroccan mosquito atlas [4] based on bibliographic work; generating presence data.

In this study, we relied on three species of mosquitoes: *Culex pipiens (Cx. pipiens), Culex theileri (Cx. theileri),* and *Culiseta longiareolata (Cs. longiareolata)*.

Table 1. Number of points of presence and absence by species (after data preprocessing)

Species	Presence	Absence
Cx. pipiens	255	105
Cx. theileri	114	187
Cs. longiareolata	133	127

A pre-processing phase of the dataset was performed including cleaning and transformation of the data: the cleaning step consists in removing the missing values and the transformation step consists in formulating all the occurrence data of the dataset in binary form (Table 1).

3.2 Performance measures

In machine learning, the evaluation of the performance of models is an important but difficult step. Several evaluation criteria were therefore calculated to evaluate the models: accuracy, Matthews Correlation Coefficient (MCC), and area under the curve ROC (AUC). The accuracy represents the ratio of correctly predicted instances to all instances in the dataset. The area under the curve ROC (AUC) is a synthetic index frequently used in predictive modeling [29]; it corresponds to the probability that a positive event (presence of the species) is classified as positive by the test over several possible thresholds. The Matthews Correlation Coefficient (MCC) is a measure often used in binary classifications to evaluate the performance of classification models; its utility has been demonstrated in various studies that concern the evaluation of predictive models [30].

3.3 Methodology

The modeling procedure implemented consists of several steps:

Step 1: Data balancing

This is a technique that aims to give equal weight to the presence and absence classes for each mosquito species. The entire minority class is used in the modeling with an equal sample size, randomly selected from the majority class [2].

Step 2: Feature selection

The number of features selected differs from one technique to another. The number of features selected by the RFECV technique depends on the algorithm, while the M10 technique selects a fixed number of features. For the rest of the techniques a feature ranking list was determined, and the top ranked features were selected using five thresholds: 10%, 20%, 40%, 50%, and log2 (number of features) [31].

Step 3: Training and evaluation of the models

The six selected learning algorithms, with the different subsets of the obtained features as well as the set of all features, were trained using a 5-fold cross-validation and evaluated by three performance criteria: accuracy, MCC coefficient, and AUC score.

Step 4: Model comparison using the SK test

The SK test [32] allows comparing several models in terms of performance in order to conclude the existence of a significant difference between them [31, 33-35]. The comparison was performed using the MCC criterion.

Step 5: Rating of the models by the Borda voting system

This voting system [36] is applied to the models belonging to the best cluster of the SK test; the three performance criteria (accuracy, MCC, and AUC) were used for this step.

Step 6: Selection of the best-scored models

Table 2. Table of abbreviations of the selection techniques used

Selection technique	Abbreviation
All features	Com
RFECV + algorithm XGBoost	RC-XG
RFECV + algorithm Logistic Regression	RC-LG
RFECV+ algorithm Random Forest	RC-RF
RFECV+ algorithm Gradient Boosting	RC-GB
ReliefF	R
Info Gain	IG
Correlation Pearson	CP
Correlation quantitative-qualitative	CQ
Feature selection according to importance	M10

This selection is based on the application of K-means algorithm with preserving the homogeneity of the identified group.

For clarity, the following abbreviations (Table 2) were used in the rest of the paper.

Feature selection and model computation were performed using the ITMO-FS and Scikit-Learn python libraries [37], while the SK statistical test was performed using R software [38].

4. RESULTS AND DISCUSSIONS

As the processed dataset contains 225 features, the subsets selected by the different techniques are summarized in Table 3.

Table 3. Number of features selected by technique

Selection technique	Number of features
RFECV	Between 10 and 103 features
M10	10 features
	Log_2 (225) (8 features)
ReliefF/Info-Gain/ Correlation	10% (23 features)
Pearson/ Correlation	20% (45 features)
quantitative-qualitative	40% (90 features)
	50% (113 features)

For the modeling of each mosquito species, a total of 156 models were computed using six learning algorithms and 26 groups of selected variables, also including the set of all features.

The SK test identified two clusters with 72 models in the best cluster for *Cs. Longiareolata* (Table 4), with a strong presence of the Gradient Boosting, Random Forest, and XGBoost models.

Table 4. Number of appearances of each algorithm in the best SK cluster and the selection techniques used for *Cs. Longiareolata*

Models	Selection techniques
Gradient Boosting (19)	RC-GB, R5, R4, RC-RF, RC-XG, CQ5, R2, IG2, CP5, IG5, CQ4, M10, CP1, IG1, IG-LOG, CQ1, CQ2, IG4, Com
Random Forest (19)	RC-GB, RC-XG, R5, RC-RF, M10, CQ5, R2, CP4, R1, R4, IG5, IG1, IG-LOG, CP5, CQ4, CQ1, CP-LOG, RC-LG, Com
XGBoost (17)	RC-GB, M10, Com, RC-XG, RC-RF, CQ4, R4, CP5, CQ5, R5, CQ-LOG, IG1, CP1, IG5, CQ1, CQ2, R2
Logistic Regression (9)	CQ1, R5, RC-RF, M10, R2, Com, R4, RC-XG, CQ-LOG
KNN (6)	CP2, CP5, CP4, CQ5, CQ4, CQ-LOG
Gaussian Naïve Bayes (2)	CP2, CP-LOG

The different models were divided into two clusters, with 79 models for the best cluster for *Cx. pipiens* (SK test result). Gradient Boosting, Random Forest, and XGBoost models are more frequent than the other models, as it is shown in Table 5, and we also observe the presence of the RFECV selection technique in the different models.

Table 5. Number of appearances of each algorithm in the best SK cluster and the selection techniques used for *Cx. pipiens*

Models	Selection techniques
Gradient Boosting (24)	RC-GB, RC-RF, CQ1, CQ4, Com, M10, CP5, CP1, CP-LOG, R2, R4, CQ2, RC-LG, CP4, CQ5, R5, CQ-LOG, R1, IG1, IG5, RC-S1, IG2, CP2, IG-LOG
Random Forest (21)	RC-RF, RC-GB, R5, Com, R2, IG5, CQ4, CQ-LOG, RC-S2, CP4, R4, IG2, CQ5, CQ2, M10, CP1, CP2, CP5, CP-LOG, RC-XG, R1
XGBoost (21)	RC-GB, R5, CP4, RC-LG, R4, CP5, IG4, R2, IG5, RC-S3, CQ1, Com, CQ2, RC-XG, IG1, CP2, CP-LOG, CQ5, M10, CQ-LOG, CP1
KNN (10)	IG2, RC-LG, RC-RF, CP-LOG, RC-GB, IG5, IG4, CQ1, M10, CQ4
Logistic Regression (2)	RC-LG, CQ2
Gaussian Naïve Bayes (1)	RC-RF

The SK test identified two clusters with 68 models in the best cluster for *Cx. theileri*, with a strong presence of Gaussian Naïve Bayes models in the best cluster, as it is shown in Table 6, and we also observe the presence of the RFECV selection technique in the different models.

Table 6. Number of appearances of each algorithm in the best SK cluster and the selection techniques used for *Cx. theileri*

Models	Selection techniques
Gaussian Naïve Bayes (17)	CQ5, CP2, RC-XG, CQ4, RC-RF, CP4, CQ1, CQ2, M10, CP5, RC-LG, CP1, IG5, CP-LOG, CQ-LOG, R4, Com
Logistic Regression (13)	Com, R4, R5, RC-XG, CP4, CP5, RC-S4, RC-RF, CP1, CP-LOG, CQ-LOG, M10, RC-LG
Gradient Boosting (11)	RC-XG, R4, RC-GB, CQ4, RC-RF, CQ5, CP1, IG5, R5, Com, M10
KNN (11)	CQ5, IG1, CP2, RC-S4, M10, CQ2, CP4, Com, IG2, RC-RF, R5
Random Forest (10)	CP1, CP2, CQ4, CQ5, RC-XG, CQ1, M10, RC-GB, RC-RF, Com
XGBoost (6)	RC-XG, R5, R4, RC-RF, Com, M10

Subsequently, the Borda voting system was applied to the models of the best cluster obtained by the SK test. This system gives scores to the models according to three performance measures (accuracy, MCC coefficient, and AUC) in order to select the best-performing models.

The selection of the best-scored models can be achieved by fixing a threshold or by using a partition procedure that puts the data into homogeneous groups. The selection threshold varies in the literature from one research to another: Idri et al. [39] chose to consider only the three best-scored models, while Hosni et al. [40] decided to select the best-scored model, whereas Benhar et al. [31] opted for the threshold of the ten best-scored models. These thresholds may be acceptable, but it is difficult to argue with such choices. Moreover, the systematic use of a predefined threshold also has the disadvantage that it may not produce a homogeneous group. Indeed, it cannot be excluded that in some cases the best-scored models present some scores that are quite different, which significantly reduces the homogeneity of this selected group.

K-means is a clustering algorithm designed to create homogeneous groups, the homogeneity increases with the number of groups. The procedure proposed in this paper consisted in applying K-means for several numbers of groups by ensuring beforehand a suitable inertia rate (exceeding 85%). The best-scored group of models is the group found by two successive applications of K-means with the same best group. Such a result guarantees the homogenization stability of the selected group, and this property is widely sought. This scenario of maintaining the selected group may not exist, and in this case, K-means will be applied for several numbers of groups until the obtained group contains the best-scored model. When a group of best-scored models is identified by successive applications of K-means, it is possible to assess its homogeneity by calculating the ratio of the variance of this group to that of the models of the first cluster of the SK test. This ratio must be lower than a certain threshold, classically 0.05.

The application of K-means on the different scores obtained by Borda gives the following results:

- Concerning *Cx. pipiens* (Figure 1), the highest-scoring group contains nine models. The performance of these models varies between 0.36 and 0.39 for the MCC coefficient, between 0.73 and 0.76 for the AUC, and between 0.67 and 0.70 for the accuracy. The homogeneity of this group is confirmed by the ratio of variances which is equal to 0.042.
- Concerning *Cs. longiareolata* (Figure 2), the highest-scoring group contains one model. This model achieved a performance equal to 0.51 for the MCC coefficient, 0.8 for the AUC, and 0.75 for the accuracy with 10 selected features. It is possible to justify some heterogeneity between this model and the second-highest-scoring model as long as the ratio of variances exceeds 0.05.
- For *Cx. theileri*, two groups are identified by successive applications of K-means; the numbers of these groups are 9 and 7, respectively. The second group (Figure 3) was selected based on its homogeneity. Indeed, the ratio of variances of the second group selected is less than 0.05 while that of the first exceeds 0.11. The performance of the models in this group varied between 0.38 and 0.43 for the MCC coefficient, between 0.70 and 0.75 for the AUC, and between 0.69 and 0.73 for the accuracy.

From the different results obtained, several remarks can be given:

- Based on the top-scored groups selected by K-means, it can be seen that most of the subsets on which the best models are based; are obtained mainly by the RFECV selection technique; this agrees with the conclusion of Misra et al. [11].
- The Gradient Boosting, Random Forest, and XGBoost algorithms were proven to be very powerful as all the best-performing models are trained using these algorithms.
- The absence of the complete models in the best K-means groups for both *Cx. pipiens* and *Cs. longiareolata* (Figures 1 and 2); can be inferred that the performance of the selected features outperformed the complete set in the modeling; hence comes the importance of feature selection.
- The presence of the M10 selection in the top-scored group for *Cx. theileri* (Figure 3); means that this method competes in terms of performance with the other selection techniques tested. The result obtained for this species confirms the

approach adopted in the research works [2, 3, 6] on the threshold of 10 features.

- The best models can be used to determine common ecological factors, specific factors, and absent factors for each mosquito species.

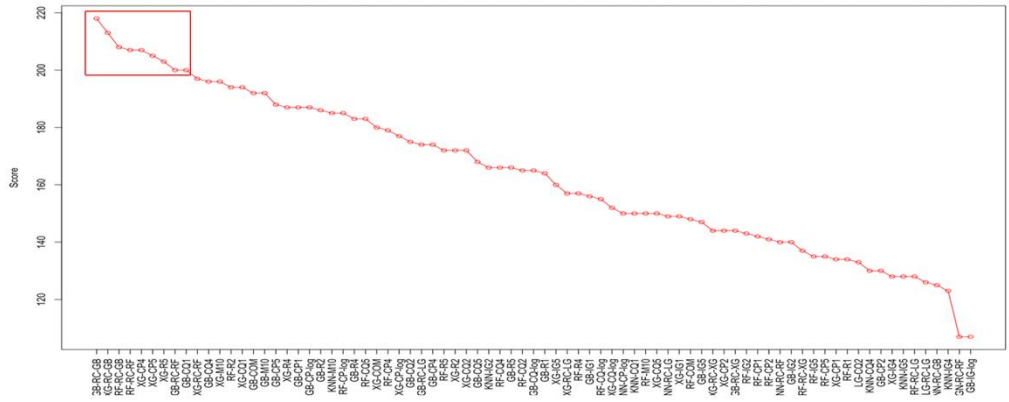

Figure 1. Graph of the scores obtained by the Borda method sorted by descending order for *Cx. pipiens*

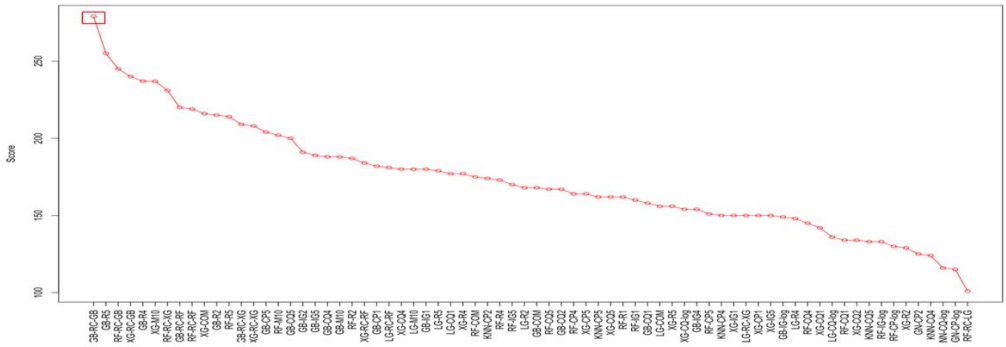

Figure 2. Graph of the scores obtained by the Borda method sorted by descending order for *Cs. Longiareolata*

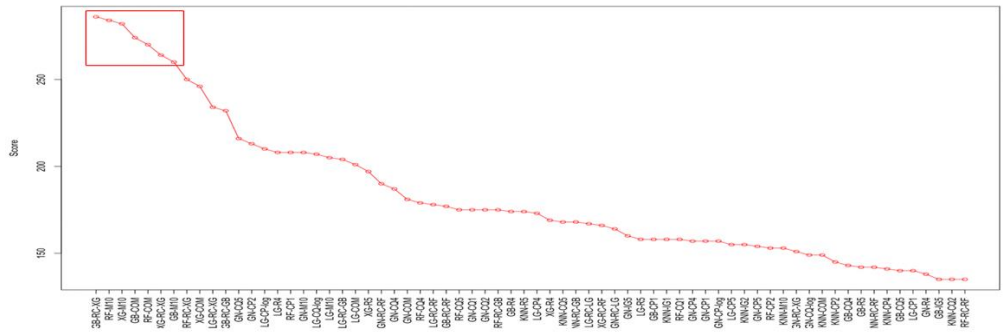

Figure 3. Graph of the scores obtained by the Borda method sorted by descending order for *Cx. theileri*

5. CONCLUSIONS

Feature selection is one of the most important data preprocessing steps. It aims to significantly improve the modeling performance by keeping only relevant features.

In this work, we compared the efficiency of the RFECV technique against a group of feature selection techniques on mosquito distribution data using a set of learning algorithms. The different models obtained following the application of the feature selection techniques and the learning algorithms were evaluated using three performance criteria: accuracy, MCC, and AUC. The selection of the best models from a large number of models is a complex step; thus the use of model comparison methods is necessary. Three steps were followed to do this comparison: the first one is the SK test which allowed us to make a first comparison of the models using the MCC criterion. Based on the results of this first comparison, the Borda voting system scored the models of the best group selected by the SK test. Then, a partitioning technique based on the K-means algorithm and the variance ratio was used to identify a group of the best models.

The results obtained showed the efficiency of the RFECV technique in modeling using the selected features. The models obtained using this recursive method outperformed both the complete models and the models obtained using the other selection techniques.

The approach adopted in this study can be recommended for modeling a dataset with a large number of features. The group

of models obtained from this approach should be significantly better than the complete model, which justifies the importance of feature selection techniques.

The dataset is rich in its 225 characteristics; the use of other selection techniques in future work should only improve the modeling quality.

ACKNOWLEDGMENT

This work is supported by the National Center for Scientific and Technical Research (CNRST), Morocco.

REFERENCES

[1] Sen, P.C., Hajra, M., Ghosh, M. (2020). Supervised classification algorithms in machine learning: A survey and review. In Emerging technology in modelling and graphics, pp. 99-111. https://doi.org/10.1007/978-981-13-7403-6_11

[2] Cianci, D., Hartemink, N., Ibáñez-Justicia, A. (2015). Modelling the potential spatial distribution of mosquito species using three different techniques. International Journal of Health Geographics, 14(1): 1-10. https://doi.org/10.1186/s12942-015-0001-0

[3] Diarra, M., Fall, M., Fall, A.G., Diop, A., Lancelot, R., Seck, M.T. (2018). Spatial distribution modelling of Culicoides (Diptera: Ceratopogonidae) biting midges, potential vectors of African horse sickness and bluetongue viruses in Senegal. Parasites & Vectors, 11(1): 1-15. https://doi.org/10.1186/s13071-018-2920-7

[4] Trari, B., Dakki, M. (2017). Atlas des moustiques (Diptera Culicidae) du Maroc. Travaux de l'Institut Scientifique, Rabat, Série Zoologie, (51): 128.

[5] Abdelkrim, O., Samia, B., Said, Z., Souad, L. (2021). Modeling and mapping the habitat suitability and the potential distribution of Arboviruses vectors in Morocco. Parasite, 28: 37. https://doi.org/10.1051/parasite/2021030

[6] Ibañez-Justicia, A., Cianci, D. (2015). Modelling the spatial distribution of the nuisance mosquito species Anopheles plumbeus (Diptera: Culicidae) in the Netherlands. Parasites & Vectors, 8(1): 1-9. https://doi.org/10.1186/s13071-015-0865-7

[7] Früh, L., Kampen, H., Kerkow, A., Schaub, G.A., Walther, D., Wieland, R. (2018). Modelling the potential distribution of an invasive mosquito species: comparative evaluation of four machine learning methods and their combinations. Ecological Modelling, 388: 136-144. https://doi.org/10.1016/J.ECOLMODEL.2018.08.011

[8] Kumar, V., Minz, S. (2014). Feature selection: A literature review. SmartCR, 4(3): 211-229. https://doi.org/10.6029/smartcr.2014.03.007

[9] Venkatesh, B., Anuradha, J. (2019). A review of feature selection and its methods. Cybernetics and Information Technologies, 19(1): 3-26. https://doi.org/10.2478/cait-2019-0001

[10] Chandrashekar, G., Sahin, F. (2014). A survey on feature selection methods. Computers & Electrical Engineering, 40(1): 16-28. https://doi.org/10.1016/j.compeleceng.2013.11.024

[11] Misra, P., Yadav, A.S. (2020). Improving the classification accuracy using recursive feature elimination with cross-validation. International Journal on Emerging Technologies, 11(3): 659-665.

[12] Wibawa, M.S., Novianti, K.D.P. (2017). Reduksi fitur untuk optimalisasi klasifikasi tumor payudara berdasarkan data citra FNA. E-Proceedings KNS&I STIKOM Bali, 73-78.

[13] Kononenko, I., Robnik-Sikonja, M., Pompe, U. (1996). ReliefF for estimation and discretization of attributes in classification, regression, and ILP problems. Artificial intelligence: methodology, systems, applications, 31-40.

[14] Kira, K., Rendell, L.A. (1992). A practical approach to feature selection. In Machine learning proceedings 1992, pp. 249-256. https://doi.org/10.1016/B978-1-55860-247-2.50037-1

[15] Robnik-Šikonja, M., Kononenko, I. (2003). Theoretical and empirical analysis of ReliefF and RReliefF. Machine Learning, 53(1): 23-69. https://doi.org/10.1023/A:1025667309714

[16] Yildirim, P. (2015). Filter based feature selection methods for prediction of risks in hepatitis disease. International Journal of Machine Learning and Computing, 5(4): 258. https://doi.org/10.7763/ijmlc.2015.v5.517

[17] Jain, D., Singh, V. (2018). Feature selection and classification systems for chronic disease prediction: A review. Egyptian Informatics Journal, 19(3): 179-189. https://doi.org/10.1016/J.EIJ.2018.03.002

[18] Eid, H.F., Hassanien, A.E., Kim, T.H., Banerjee, S. (2013). Linear correlation-based feature selection for network intrusion detection model. In International Conference on Security of Information and Communication Networks, pp. 240-248. https://doi.org/10.1007/978-3-642-40597-6_21

[19] Sanharawi, M.E., Naudet, F. (2013). Comprendre la régression logistique. Journal Francais D Ophtalmologie, 36: 710-715. https://doi.org/10.1016/J.JFO.2013.05.008

[20] Kamel, H., Abdulah, D., Al-Tuwaijari, J.M. (2019). Cancer classification using gaussian naive bayes algorithm. In 2019 International Engineering Conference (IEC), pp. 165-170. https://doi.org/10.1109/IEC47844.2019.8950650

[21] Cover, T.M., Hart, P.E. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13: 21-27. https://doi.org/ 10.1109/TIT.1967.1053964

[22] Breiman, L. (2001). Random forests. Machine Learning, 45(1): 5-32. https://doi.org/10.1023/A:1010933404324

[23] Chickaramanna, S.G., Veerabhadrappa, S.T., Shivakumaraswamy, P.M., Sheela, S.N., Keerthana, S.K., Likith, U., Swaroop, L., Meghana, V. (2022). Classification of arrhythmia using machine learning algorithm. Revue d'Intelligence Artificielle, 36(4): 529-534. https://doi.org/10.18280/ria.360403

[24] Friedman, J.H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 1189-1232. https://doi.org/ 10.1214/AOS/1013203451

[25] Lusa, L. (2017). Gradient boosting for high-dimensional prediction of rare events. Computational Statistics & Data Analysis, 113: 19-37. https://doi.org/10.1016/j.csda.2016.07.016

[26] Chen, T., Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp. 785-794.

https://doi.org/10.1145/2939672.2939785
[27] Miriyala, N.P., Kottapalli, R.L., Miriyala, G.P., Lorenzini, G., Ganteda, C., Bhogapurapu, V.A. (2022). Diagnostic analysis of diabetes mellitus using machine learning approach. Revue d'Intelligence Artificielle, 36(3): 347-352. https://doi.org/10.18280/ria.360301
[28] Bennouna, A. (2019). Chorologie, écologie et caractérisation du virome de populations de moustiques au Maroc: Exemple de Culex pipiens vecteur majeur d'arbovirus. Doctorat de l'Institut Agronomique et Vétérinaire Hassan II, Sciences vétérinaires. Soutenue le12 septembre 2019.
[29] Yang, S., Berdine, G.G. (2017). The receiver operating characteristic (ROC) curve. The Southwest Respiratory and Critical Care Chronicles, 5: 34-36. https://doi.org/10.12746/swrccc.v5i19.391
[30] Chicco, D., Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics, 21(1): 1-13. https://doi.org/ 10.1186/s12864-019-6413-7
[31] Benhar, H., Idri, A., Hosni, M. (2020). Impact of threshold values for filter-based univariate feature selection in heart disease classification. In HEALTHINF, pp. 391-398. https://doi.org/10.5220/0008947403910398
[32] Scott, A.J., Knott, M. (1974). A cluster analysis method for grouping means in the analysis of variance. Biometrics, 507-512. https://doi.org/ 10.2307/2529204
[33] Azzeh, M., Nassif, A.B., Banitaan, S. (2017). Comparative analysis of soft computing techniques for predicting software effort based use case points. IET Softw., 12: 19-29. http://doi.org/10.1049/iet-sen.2016.0322
[34] Ghotra, B., McIntosh, S., Hassan, A. (2015). Revisiting the impact of classification techniques on the performance of defect prediction models. 2015 IEEE/ACM 37th IEEE International Conference on Software Engineering, pp. 789-800. http://doi.org/10.1109/ICSE.2015.91
[35] Chlioui, I., Abnane, I., Idri, A. (2020). Comparing statistical and machine learning imputation techniques in breast cancer classification. In International Conference on Computational Science and Its Applications, pp. 61-76. https://doi.org/ 10.1007/978-3-030-58811-3_5
[36] Van Erp, M., Vuurpijl, L., Schomaker, L. (2002). An overview and comparison of voting methods for pattern recognition. In Proceedings Eighth International Workshop on Frontiers in Handwriting Recognition, pp. 195-200. https://doi.org/10.1109/IWFHR.2002.1030908
[37] Pilnenskiy, N., Smetannikov, I. (2020). Feature selection algorithms as one of the python data analytical tools. Future Internet, 12(3): 54. https://doi.org/10.3390/fi12030054
[38] Jelihovschi, E.G., Faria, J.C., Allaman, I.B. (2014). ScottKnott: a package for performing the Scott-Knott clustering algorithm in R. TEMA (São Carlos), 15: 3-17. https://doi.org/ 10.5540/TEMA.2014.015.01.0003
[39] Idri, A., Bouchra, E.O., Hosni, M., Abnane, I. (2020). Assessing the impact of parameters tuning in ensemble based breast Cancer classification. Health and Technology, 10(5): 1239-1255. https://doi.org/10.1007/s12553-020-00453-2
[40] Hosni, M., Idri, A. (2017). Software effort estimation using classical analogy ensembles based on random subspace. In Proceedings of the Symposium on Applied Computing, pp. 1251-1258. https://doi.org/10.1145/3019612.3019784