Revue d'Intelligence Artificielle
Vol. 38, No. 4, August, 2024, pp. 1069-1076
Journal homepage: http://iieta.org/journals/ria

A Procedure to Improve Binary Classification Models and Categorize Features: The Case of the Distribution of Three Mosquito Species in Morocco

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

[1] Advanced Digital Enterprise Modeling and Information Retrieval Laboratory, ENSIAS, Mohammed V University in Rabat, Rabat 10000, Morocco
[2] Smart Systems Laboratory, ENSIAS, Mohammed V University in Rabat, Rabat 10000, Morocco
[3] CIRAD, UMR ASTRE, Montpellier F-34398, France
[4] ASTRE, University of Montpellier, CIRAD, INRAE, Montpellier 34398, France
[5] Formerly Microbiology, Immunology and Contagious Diseases Unit, Agronomic and Veterinary Institute Hassan II, Rabat BP 6202, 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.380402

Received: 16 November 2023
Revised: 2 June 2024
Accepted: 23 July 2024
Available online: 23 August 2024

Keywords:
feature selection, improving performance, multiple solutions, categorization of features, mosquito

ABSTRACT

Modeling biological datasets represents an essential step in processing and exploiting biological information. Selecting features and improving modeling quality are critical in building a high-performance predictive model. In this article, we have presented and applied a novel approach to select features and to improve the modeling quality using the presence/absence data of three mosquito species in Morocco. This approach uses a recursive search of feature subsets conditioned on improving the modeling quality compared to an initially chosen solution. It has led to a significant improvement in the modeling quality compared to another study carried out on the same dataset, where the accuracy of the models improved with a range varying between 0.062 and 0.198. The relevance of this approach also extends to the search for solutions that achieve the same performance with different subsets, known as multiple solutions. These solutions demonstrate that various combinations of explanatory features can explain the target feature, leading to categorizing them according to their impact on the modeling. This work has provided a good explanation of the distribution of mosquito species thanks to the improved modeling quality, opening up the possibility of having relevant solutions and discovering new explanatory modes for the features.

1. INTRODUCTION

With the increasing quantity of data available in the biological field, the tools needed to process it have become more and more necessary. Today, computer science offers many tools and techniques that are indispensable for the analysis and interpretation of biological data containing a large number of features. These tools include machine learning algorithms, processing techniques, modeling software, data visualization tools, and high-performance computing infrastructures for data storage and processing.

Machine learning is a computing technique increasingly used in analyzing biological data. This technique has made it possible to develop algorithms that learn from biological data, enabling the identification of patterns and associations between different environmental features. Many studies have demonstrated the benefits of using machine learning to analyze biological data, for example:

- González Jiménez et al. [1] proposed an approach for predicting the age of mosquitoes and identifying their species using machine learning algorithms. This approach has proved effective in terms of speed, cost, and accuracy compared with traditional methods of determining the age and species of mosquitoes, which has helped to improve malaria control and management strategies.
- Yang et al. [2] presented four uses of machine learning in the domain of DNA sequences. In the case of alignment, the Genetic Algorithm stands out for its computational speed, its efficiency, which is resistant to the length and number of sequences, and its relevance and accuracy.
- Abhari et al. [3] examined various applications of machine learning methods in managing type 2 diabetes. They underscored the efficiency and accuracy of machine learning algorithms, such as Support Vector Machine and Naive Bayesian, in classifying diabetes. These methods can contribute to developing patient treatment plans and improving disease management.

The quality of modeling occupies a significant place in the predictive analysis field. Indeed, a high-quality model is characterized by its explanatory power. In this context,

selecting features and improving modeling quality are essential steps in building a high-performance predictive model, especially when the dataset to be processed contains a large number of features.

The study of mosquitoes is an important field of biological research in Morocco due to the great diversity of these species [4] and their role in the transmission of pathogens [5, 6].

Douider et al. [7] modeled the distribution of three mosquito species using 225 environmental factors divided into 11 groups from online ecological datasets. The diversity of environmental factors available in the dataset and the availability of presence and absence records for each species constitute the interest of the study [7] compared to other research [5, 8]. The modeling phase undertaken in the study [7] resulted in the use of six learning algorithms and a group of feature selection techniques, producing a set of models for each species. By comparing the models using a set of comparison techniques and a group of quality criteria, it was possible to select a set of the most-performing models. The quality of these models ranged from 0.67 to 0.75 for accuracy and from 0.36 to 0.51 for MCC. These models exhibit acceptable quality; they have improved on the models that use all the features in the dataset. However, there is still considerable opportunity for improvement in this modeling since there is a range for enhancement.

However, improving the modeling quality of mosquito data is very important for several reasons. Firstly, a good quality model can help identify the favorable and unfavorable factors influencing mosquito distribution, which can help prevent the spread of vector-borne diseases [9]. In addition, a good quality model can be used to evaluate the effectiveness of mosquito control interventions, which can help optimize the use of public health resources and improve the efficiency of control programs [5, 10].

This study introduces a new procedure for selecting features and improving modeling quality. This procedure is based on a recursive search of feature subsets, conditional on improving modeling quality, starting from a chosen initial solution. The effectiveness of this procedure was evaluated using data on the distribution of three mosquito species [7]. Improving the modeling quality of this data allows for a more precise explanation of the target feature based on environmental features. This enhancement would provide a better understanding of the distribution of each species, thus contributing to a richer knowledge of mosquitoes. The learning algorithms chosen to implement this method are Gradient Boosting, XGBoost, and Random Forest, which are highly distinguished in the research [7].

Once a model with satisfactory quality has been obtained, it is interesting to search for the presence of other feature subsets with the same level of quality. In general, a better solution to a binary classification model is not necessarily unique, and the search for other feature subsets of similar modeling quality can only be of great use. The presence of such solutions means that the explanation of the target feature by the explanatory features is not unique; each solution can illustrate a particular scenario of the presence or absence of mosquitoes. This can only enrich scientific knowledge of this phenomenon.

The remainder of this paper is organized as follows: Section 2 introduces a set of techniques for feature selection and modeling improvement. Section 3 outlines the proposed methodology. In Section 4, experimental results and performance analysis are presented. Finally, Section 5 provides the conclusion of the study.

2. FEATURE SELECTION AND IMPROVEMENT OF MODELING QUALITY METHODS

2.1 Feature selection methods

Feature selection represents one of the most frequently employed approaches for dimensionality reduction in data analysis [11]. It aims to build an improved model by selecting a subset of features from the original set according to the meaning and relevance of those features [12]. Feature selection techniques can be categorized into several types:

- Filter: In this category, features are selected independently of the learning algorithm using statistical measures. Some of the commonly used statistical metrics for calculating feature importance include Information Gain [13], Chi-Square test [14], ReliefF [15], and Correlation Coefficient [16].
- Wrapper: This method selects a subset of features based on their predictive capacity against a specific learning algorithm. It uses a search algorithm to explore the feature space and identify the best subset of features that yield optimal performance. Among the search algorithms, we cite Recursive Feature Elimination with Cross-Validation [7, 17] and Sequential Feature Selection algorithms (Backward and Forward) [12, 18, 19].
- Embedded: The process of selecting features in this type of method is integrated into the learning algorithm and depends on its properties. Some of the most popular classification algorithms for integrated feature selection methods are Support Vector Machine, Artificial Neural Network, and Decision Tree methods [20].

Another method that has recently emerged is ensemble feature selection [21]. This method combines the outputs of a group of selection techniques, such as ReliefF and Pearson's Correlation Coefficient, and then produces an aggregated result [22, 23].

2.2 Combination of selection methods

In the field of ensemble feature selection, the combination of outcomes from a group of selection methods yields an aggregated result [24]. These combination procedures are categorized based on the type of result obtained by each method:

- If the selection methods produce scores for each feature, the most prevalent combination procedure involves aggregating the scores, which can be achieved using techniques such as mean, median, or maximum [22, 24].
- If the methods produce subsets of features, the common typical combination procedure is intersection, which allows only the features common to all the methods to be selected, and union, which allows the features selected by all methods [25, 26].

2.3 Feature selection with depth-first search (FSDFS)

FSDFS is a proposed wrapper method for selecting features and improving modeling quality using graph theory. This method can produce very interesting results, yet it is

based on a simple procedure: the quality of a model M with k features can be improved by either removing an existing feature or adding another (Figure 1).

This idea can be optimally exploited using the depth-first search algorithm, which eliminates to avoid re-exploring processed subsets.

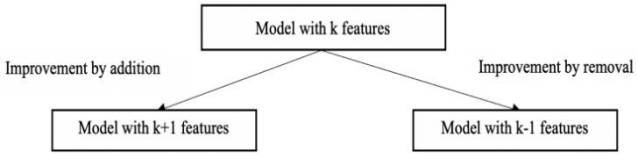

Figure 1. Procedure for improving the quality of a model with k features

Let X be the set of n features available in the dataset to be processed and Θ the chosen quality criterion. The objective is to find improved solutions according to Θ from the initial subset M.

The process of searching for improved solutions consists of adding and removing features. Three functions were developed to simplify the programming of this process:

Algorithm 1: FSDFS

Input: $X, M,$ *Learning algorithm,* $\Theta, \Theta(M)$
Declaration of empty lists:$Explored\_models,\ Quality,\ Improved\_models$
function Improved $(X, \Theta, FS, \Theta(M))$
  for *i=1 to* n ***do***
    $Neighbor \leftarrow FS$
    $Neighbor[i] \leftarrow not\ FS[i]$
    if $Neighbor$ *does not exist in* $Explored\_models$ ***then***
      add $Neighbor$ *to* $Explored\_models$
      $X\_Neighbor[i] \leftarrow$ ***Retransform*** $(Neighbor)$
      Calculate the performance of the $X\_Neighbor$ *subset*
      if $\Theta(M) < \Theta(X\_Neighbor)$ ***then***
        add $X\_Neighbor$ *to* $Improved\_models$
        add $\Theta(X\_Neighbor)$ *to* $Quality$
        Improved $(X, \Theta, Neighbor, \Theta(X\_Neighbor))$
      end if
    end if
  end for
end function
begin:
$FS \leftarrow$ ***Transform***(M)
add FS *to* $Explored\_models$
Improved $(X, \Theta, FS, \Theta(M))$
end
Output: $Improved\_models, Quality$

- The '***Transform***' function converts the subset M of k features into a binary list of size n: the 'True' values in the binary list correspond to the features in the subset M.
- The '***Retransform***' function is the reciprocal of the first function, it transforms a binary list FS of size n with k 'True' elements into a subset M of k features.
- The '***Improved***' function presents a recursive form for searching the improved solutions of an initial subset M. It takes as inputs: the set X, the quality criterion Θ, the binary list FS of the features of the initial subset M and its quality level $\Theta(M)$. The function begins by determining new subsets that differ from the initial subset M by a single feature and evaluating their quality. Each processed subset will be stored in the '*Explored_models*' list to avoid reusing previously processed subsets. When the quality of a subset exceeds that of M, this quality will be added to a '*Quality*' list and the subset of features to an '*Improved_models*' list. In this case, the process of finding improved solutions will repeat, but this time with the new improved subset as the solution to be processed. This repetition will broaden the '*Explored_models*' list, and the search process for improved solutions will continue. This process will stop when all subsets that differ from M by one feature have been processed. This condition limits the possibilities of improving the initial solution, but the algorithm as it is designed can process up to $2^n - 1$ subsets of the features, which is unfeasible once n exceeds a certain threshold, and in these conditions, the execution time can be very long.

The '*Improved_models'* list obtained by FSDFS (Algorithm 1) displays a set of models that have been improved compared with the initial model M.

The FSDFS algorithm can also be used in the search for multiple solutions that achieve similar performance. It allows the exploration of the solution space and the identification of different combinations of features that may be effective in modeling a dataset.

The quality of modeling a dataset is intricately associated with the features utilized during the modeling process. Using all existing features does not necessarily guarantee optimal quality, particularly in the presence of redundant and insignificant features. Moreover, an increased number of features makes the modeling and interpretation phase more complex. Feature selection methods aim to choose a subset of features that yield the best modeling results using a variety of approaches. Some methods use statistical tests, others evaluate the predictive power of features using a chosen learning algorithm, while others apply the principle of addition and removal. Focusing on a relevant subset of features improves modeling quality and facilitates interpretation.

3. METHODOLOGY

Selecting features and improving the modeling quality are essential steps in the machine learning application. These procedures play an important role in improving the performance and efficiency of machine learning models.

The methodology proposed in this article was evaluated using a dataset on the distribution of three mosquito species in Morocco [7]. The dataset was collected from 366 sites and included the target feature (the presence or absence of mosquitoes) and 225 environmental features.

3.1 Selecting features and improving modeling

The proposed methodology for selecting and improving the modeling quality of a dataset is presented in Figure 2 and consists of five main steps:

- Step 1: Data pre-processing

This step is a preliminary phase in the data modeling

process, which aims to transform raw data into a usable form. The pre-processing operations used in this study are data cleaning, data transformation, and data balancing [27, 28].

- Step 2: Selection of an initial solution

The principle of this step is to select an initial model that is better than the model with all the features. This selection can be realized using different feature selection techniques.

- Step 3: Application of the FSDFS algorithm

This algorithm starts its search from an initial solution and generates new subsets by adding and removing features from X. Given the nature of the algorithm, if n exceeds a certain threshold, the execution time will be very long. A time threshold can be fixed to interrupt the search process if the algorithm continues to run.

- Step 4: Evaluation of results

Among the results of the FSDFS algorithm, there is a list of subsets of features, all of which have a better quality than the initial subset. When the highest quality of these subsets is satisfactory, the procedure stops. If not, a group of the best subsets is selected to go on to step 5.

- Step 5: Intersection of solutions

The intersection technique proposes a new initial solution for the FSDFS algorithm by combining the best subsets obtained. It is possible to repeat this process several times until the quality obtained is satisfactory.

The modeling quality is satisfactory when it exceeds a threshold considered suitable. However, a quality level that is not improved after using a new initial solution may prove acceptable, and the possibility of repeating the process using another solution may be worth considering.

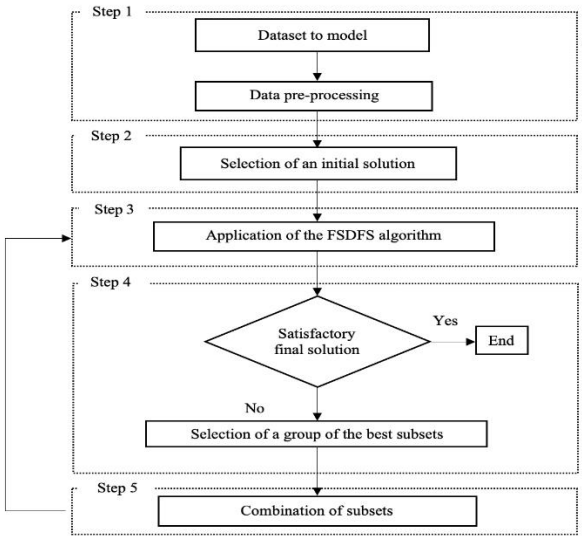

Figure 2. Methodology proposed

3.2 Multiple solutions search

A multiple solution in terms of performance and quality refers to several solutions that achieve equal performance with different subsets of features, while using the same learning algorithm. This means that different combinations of features can achieve the same performance.

The methodology used to investigate the existence of multiple solutions consists of applying the FSDFS algorithm to the best model obtained by modifying the subset processing condition to display feature subsets whose quality equals that of the best model.

Multiple solutions present groups of features that ensure the same level of explanation of the target feature. The occurrence of each explanatory feature in these groups is an interesting indicator, as the more a feature is present in the groups, the more it contributes to explaining the target feature. Conversely, the absence of a feature in all groups indicates that it has no contribution.

4. RESULTS

This section evaluates the results of the proposed methodology for modeling the distribution of three mosquito species. The performance criteria selected for this evaluation are [29]:

- Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$
- Sensitivity = $\frac{TN}{TN+FP}$
- Specificity = $\frac{TP}{TP+FN}$
- MCC = $\frac{TP\times TN-FP\times FN}{\sqrt{(TP+FP)\times(TP+FN)\times(TN+FP)\times(TN+FN)}}$

With:

TP: the number of correctly classified presence observations.

TN: the number of correctly classified absence observations.

FP: the number of absence observations classified as presence.

FN: the number of presence observations classified as absences.

The results cover both aspects of the methodology.

4.1 Selecting features and improving modeling

Tables 1, 2, and 3 describe the best results obtained for modeling three mosquito species, showing the variation in modeling quality depending on the algorithms used.

The proposed methodology requires an initial solution. To this end, the following techniques are used: The Backward selection and the combination of the best solutions obtained in the study [7]. Then, the approach for improving the quality of the model is applied to the initial solutions. The results for the different species showed an improvement after the first application of the FSDFS technique. This was illustrated by the increase in the various performance measures. The improvement in performance after the first application of FSDFS ranged from 0.11 to 0.383 for the MCC criterion, from 0.059 to 0.189 for accuracy, from 0.045 to 0.183 for sensitivity, and from 0.05 to 0.21 for specificity. However, the most significant improvement was observed in modeling the *Cx. theileri* species using the Gradient Boosting algorithm, where the MCC criterion increased from 0.277 to 0.66.

After obtaining the intermediate results through the application of FSDFS, it is possible to combine them by selecting common predictors in the best model group. These predictors can be considered important and influential for the study of the predictions of these species. The process of combining solutions and re-executing the FSDFS approach can be iterative, continuing the gradual improvement of the model until satisfactory performance is achieved. By re-applying FSDFS, we can explore other combinations of features and check if such combinations can lead to better model performance. In general, the second application of the

FSDFS can better refine the model by improving data adjustment and identifying more relevant features. This can result in a further increase in performance measures (as in the case of modeling the *Cx. pipiens* species using the XGBoost algorithm, where accuracy is increased from 0.842 to 0.876), or in a reduction in the number of features while maintaining a level of quality similar to that of the best model group (as in the case of modeling the *Cx. theileri* species using the Gradient Boosting algorithm, where the number of features is reduced from 30 to 24).

Table 1. Modeling the *Cs. longiareolata* species by: Gradient Boosting, XGBoost, Random Forest

Gradient Boosting	Number of Features	Accuracy	Sensitivity	Specificity	MCC
Backward_MCC	9	0.811	0.792	0.848	0.635
FSDFS_MCC 'Best quality'	16	0.870	0.844	0.909	0.745
Intersection of features in the group of the best solutions	13	0.818	0.807	0.845	0.645
FSDFS_MCC 'Best quality' 'Unique model'	22	0.885	0.876	0.903	0.772

XGBoost	Number of Features	Accuracy	Sensitivity	Specificity	MCC
Backward_ Accuracy	11	0.803	0.772	0.840	0.607
FSDFS_ Accuracy 'Best quality'	34	0.862	0.817	0.919	0.733
Intersection of features in the group of the best solutions	20	0.811	0.762	0.871	0.630
FSDFS_MCC 'Best quality' 'Unique model'	35	0.881	0.847	0.925	0.766

Random Forest	Number of Features	Accuracy	Sensitivity	Specificity	MCC
Backward_ Accuracy	9	0.740	0.729	0.767	0.502
FSDFS_ Accuracy 'Best quality'	25	0.830	0.810	0.861	0.665
Intersection of features in the group of the best solutions	15	0.767	0.733	0.808	0.539
FSDFS_ Accuracy 'Best quality' 'Group of models'	23	0.838	0.805	0.878	0.678

Table 2. Modeling the *Cx. theileri* species by: Gradient Boosting, XGBoost, Random Forest

Gradient Boosting	Number of Features	Accuracy	Sensitivity	Specificity	MCC
Combination of the best solutions obtained in [7]	3	0.639	0.661	0.615	0.277
FSDFS_MCC 'Best quality'	30	0.828	0.844	0.821	0.660
Intersection of features in the group of the best solutions	15	0.767	0.806	0.734	0.536
FDSFS MCC 'Best quality' 'Unique model'	24	0.837	0.876	0.805	0.679

XGBoost	Number of Features	Accuracy	Sensitivity	Specificity	MCC
Backward_ Accuracy	11	0.745	0.783	0.726	0.505
FSDFS_MCC 'Best quality'	21	0.807	0.827	0.802	0.623
Intersection of features in the group of the best solutions	15	0.750	0.770	0.745	0.508
FSDFS_MCC 'Best quality' 'Group of models'	28	0.807	0.836	0.804	0.634

Random Forest	Number of Features	Accuracy	Sensitivity	Specificity	MCC
Backward_ Accuracy	5	0.684	0.709	0.677	0.382
FSDFS_ Accuracy 'Best quality'	23	0.776	0.804	0.763	0.562
Intersection of features in the group of the best solutions	11	0.710	0.742	0.691	0.428
FSDFS_Accuracy 'Best quality' 'Group of models'	17	0.780	0.810	0.770	0.578

Table 3. Modeling the *Cx. pipiens* species by: Gradient Boosting, XGBoost, Random Forest

Gradient Boosting	Number of Features	Accuracy	Sensitivity	Specificity	MCC
Combination of the best solutions obtained in the research [7]	4	0.666	0.681	0.663	0.351
FSDFS_MCC 'Best quality'	31	0.842	0.827	0.873	0.701
Intersection of features in the group of the best solutions	27	0.814	0.799	0.846	0.649
FSDFS_MCC 'Best quality' 'Unique model'	31	0.842	0.827	0.873	0.701

XGBoost	Number of Features	Accuracy	Sensitivity	Specificity	MCC
Combination of the best solutions obtained in the research [7]	7	0.685	0.682	0.692	0.380
FSDFS_ Accuracy 'Best quality'	21	0.842	0.831	0.854	0.685
Intersection of features in the group of the best solutions	6	0.704	0.692	0.728	0.422
FSDFS_ Accuracy 'Best quality' 'Group of models'	30	0.876	0.880	0.877	0.755

Random Forest	Number of Features	Accuracy	Sensitivity	Specificity	MCC
Combination of the best solutions obtained in the research [7]	19	0.714	0.689	0.745	0.436
FSDFS_MCC 'Best quality'	33	0.804	0.805	0.810	0.613
Intersection of features in the group of the best solutions	8	0.728	0.720	0.743	0.465
FSDFS_MCC 'Best quality' 'Unique model'	28	0.833	0.816	0.859	0.674

After applying the feature selection and modeling improvement process, the final solution can take one of two forms:

- A group of models with similar performances. This was

the case in the modeling of the Cx. pipiens species using the XGBoost algorithm;

- A single model with the best performance, as in the case of the modeling of the species Cs. longiareolata using the Gradient Boosting algorithm.

4.2 Multiple solutions

The search for multiple solutions concerned the best solutions calculated by the proposed procedure. Such a search allows for identifying subsets of features with the same explanatory power. Each subset illustrates a scenario for explaining the target feature. In this study, this operation was conducted on all the best models obtained from the FSDFS (Tables 1, 2, and 3). The results indicate the presence of multiple solutions for the XGBoost model applied to the three mosquito species.

The search for multiple solutions for the Cx. theileri species revealed the existence of four models with identical performances (Figure 3). The model with 27 features distinguishes as a multiple solution and as the intersection model of the four models. It implies that these 27 features are the most important for modeling this species using the XGBoost model. Furthermore, the two models with 28 features differ only by adding a unique supplementary feature compared to the model with 27 features, while the model with 29 features represents the union model. The insertion of supplementary features does not cause any improvement or deterioration in modeling quality. They have a redundant effect on the distribution of the Cx. theileri species. All other features not present in the multiple solutions were considered features with no impact on modeling.

The distribution of the multiple solutions for the Cs. longiareolata species follows the same pattern as for the Cx. theileri species, but this time with a set of 128 models (Figure 3). The intersection of these models results in 34 principal features, which are the most important for the distribution of this species, while the union corresponds to the model with 41 features. The other models are constructed by adding several combinations of the seven supplementary features to the 34 features of the intersection model.

The search for multiple solutions for the Cx. pipiens species revealed the existence of 129 models with similar performance. The distribution pattern of these solutions differs from the other species studied (Figure 3). One of the particularities of this species is that it presents two models with the minimum number of features, which is 28. These two models share the 27 features common to all 129 models and admit two exclusive features. Furthermore, the groups of features in these two models accept the same six supplementary features. For the other multiple models, one model is made up of the union of models with 28 features and two specific supplementary features, while the rest of the models are composed of one of the models with 28 features and a combination of the six supplementary features. The union of models with 28 features does not constitute a multiple solution; it is necessary to add two specific supplementary features to this union. These two features make it possible to obtain a multiple solution including two exclusive features.

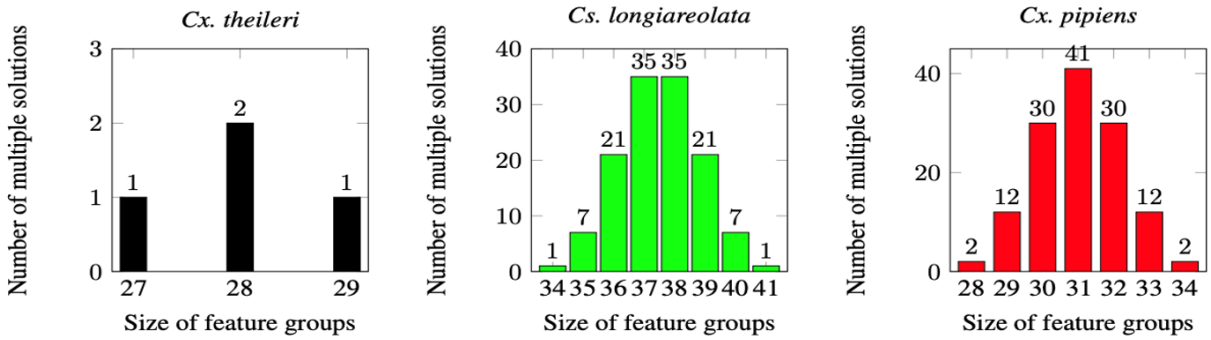

Figure 3. Distribution of multiple solutions for each mosquito species

The results obtained by identifying multiple solutions highlight the diversity of interactions between the explanatory and the target features. A categorization was determined to clarify these interactions based on the occurrence of features in the different solutions obtained. This categorization led to the definition of several concepts that can help us better understand the explanatory power of the features:

- The core represents the intersection of all of the multiple solutions, which is the set of common features. These features have a significant impact on the performance of the model and are considered to be the most important.
- The principal features belong to the union of all of the multiple solutions with the minimum number of features. They have the best explanatory capacity but are not necessarily all present in the core.
- The supplementary features have a redundant role and are characterized by their number. Any combination of these features is added to the principal features to form multiple solutions.
- The exclusive principal features, whose category has been highlighted for the Cx. pipiens species, but can be generalized. It consists of the principal features that complete the core to obtain an optimal solution. They can only be grouped in such a solution if some specific supplementary features are present.
- The features with no effect are those not included in the multiple solutions.

This categorization of explanatory features has revealed new modes of explaining the target feature: core features, supplementary features, exclusive principal features associated with specific supplementary features, and features with no effect. This information can only enrich the scientific explanation of the absence or presence of these three mosquito species. Without this categorization, a feature contributes to explaining the target feature if it is present in a multiple solution and makes no contribution if it is not.

The procedure proposed in this article has demonstrated its effectiveness on the mosquito dataset. The modeling quality

has been improved for the three mosquito species. Additionally, the search for the existence of multiple solutions has shown its value in this application. These achievements have established a robust basis for modeling these species in terms of modeling quality and the impact of explanatory features. The various results obtained enrich the field of mosquito research and data analysis. For future work, it would be interesting to confirm the categories of features identified using the interpretability criteria mentioned by Hakkoum et al. [30] and to test the ability of this proposed procedure to enhance the modeling quality of other datasets.

5. CONCLUSION

Modeling a dataset involves several steps, from data cleaning to model interpretation. Feature selection and improvement of modeling quality are essential for building a successful predictive model.

This work presents a new procedure for selecting features and improving modeling quality. It consists of two phases:

- The improvement phase (FSDFS) involves exploring features by iteratively adding or removing features from a chosen initial solution.
- The combination phase combines the features of a group of the best solutions obtained in the first phase to find a new initial solution.

The application of this procedure to the modeling of the three mosquito species led to significant performance improvements over the results obtained in [7]. The improvement gaps are remarkable for the different performance criteria. They ranged from 0.062 to 0.198 for accuracy, from 0.053 to 0.215 for sensitivity, from 0.055 to 0.21 for specificity, and from 0.129 to 0.402 for MCC. These results underline the effectiveness of the FSDFS method in improving model quality. In the improvement phase, any initial solution can be used. The solutions obtained in the combination phase proved interesting, as several improvements were obtained after processing these solutions with the FSDFS algorithm (Tables 1, 2, and 3).

The satisfactory performance of the models obtained raises questions about the existence of multiple solutions for these models. It is possible to check this by modifying the condition of processing subsets of features in the FSDFS method. By applying this process, the existence of multiple solutions was revealed only for the XGBoost model. These solutions highlighted a diversity of scenarios for the presence or absence of mosquitoes. To clarify this configuration, a categorization based on the occurrence of features in the different solutions was carried out. It has led to the discovery of new ways of explaining the features, which can only help scientific understanding of the absence and presence of these mosquito species.

ACKNOWLEDGMENT

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

REFERENCES

[1] González Jiménez, M., Babayan, S.A., Khazaeli, P., Doyle, M., Walton, F., Reddy, Glew, T., Viana, M., Ranford-Cartwright, L., Niang, A., Siria, D.J., Okumu, F.O., Diabaté, A., Ferguson, H.M., Baldini, F., Wynne, K. (2019). Prediction of mosquito species and population age structure using mid-infrared spectroscopy and supervised machine learning. Wellcome Open Research, 4: 76. https://doi.org/10.12688/wellcomeopenres.15201.3

[2] Yang, A., Zhang, W., Wang, J., Yang, K., Han, Y., Zhang, L. (2020). Review on the application of machine learning algorithms in the sequence data mining of DNA. Frontiers in Bioengineering and Biotechnology, 8: 1032. https://doi.org/10.3389/fbioe.2020.01032

[3] Abhari, S., Kalhori, S.R.N., Ebrahimi, M., Hasannejadasl, H., Garavand, A. (2019). Artificial intelligence applications in type 2 diabetes mellitus care: Focus on machine learning methods. Healthcare Informatics Research, 25(4): 248-261. https://doi.org/10.4258/hir.2019.25.4.248

[4] Aboulfadl, S., Mellouki, F., Aouinty, B., Faraj, C. (2022). Susceptibility status of Culex pipiens larvae (Diptera: Culicidae) to the main insecticides used in larval control in the regions of Rabat and Casablanca in Morocco. International Journal of Pest Management, 68(3): 267-273. https://doi.org/10.1080/09670874.2020.1818869

[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] Trari, B., Dakki, M. (2017). Atlas des Moustiques (Diptera Culicidae) du Maroc. Projet: Atlas of the mosquitoes (Diptera: Culicidae) of Morocco (North Africa). Université Mohammed V de Rabat, Institut Scientifique.

[7] Douider, M., Amrani, I., Balenghien, T., Bennouna, A., Abik, M. (2022). Impact of recursive feature elimination with cross-validation in modeling the spatial distribution of three mosquito species in Morocco. Revue d' Intelligence Artificielle, 36(6): 855-862. https://doi.org/10.18280/ria.360605

[8] Laboudi, M., Faraj, C., Rhajaoui, M., El-Aouad, R., Sadak, A., Azelmate, M. (2012). Some environmental factors associated with Anopheles labranchiae larval distribution during summer 2009, in Larache Province, Morocco. African Entomology, 20(2): 229-238. https://hdl.handle.net/10520/EJC125252

[9] 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: 1-9. https://doi.org/10.1186/s13071-015-0865-7

[10] Ciss, M., Biteye, B., Fall, A.G., Fall, M., Gahn, M.C.B., Leroux, L., Apolloni, A. (2019). Ecological niche modelling to estimate the distribution of Culicoides, potential vectors of bluetongue virus in Senegal. BMC Ecology, 19(1): 1-12. https://doi.org/10.1186/s12898-019-0261-9

[11] 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

[12] Khaire, U.M., Dhanalakshmi, R. (2019). Stability of feature selection algorithm: A review. Journal of King

Saud University - Computer and Information Sciences, 34(4): 1060-1073. https://doi.org/10.1016/j.jksuci.2019.06.012

[13] Bommert, A., Sun, X., Bischl, B., Rahnenführer, J., Lang, M. (2020). Benchmark for filter methods for feature selection in high-dimensional classification data. Computational Statistics & Data Analysis, 143: 106839. https://doi.org/10.1016/j.csda.2019.106839

[14] Gárate-Escamila, A.K., El Hassani, A.H., Andrès, E. (2020). Classification models for heart disease prediction using feature selection and PCA. Informatics in Medicine Unlocked, 19: 100330. https://doi.org/10.1016/j.imu.2020.100330

[15] Ghosh, P., Azam, S., Jonkman, M., Karim, A., Shamrat, F.J.M., Ignatious, E., Shultana, S., Beeravolu, A.R., De Boer, F. (2021). Efficient prediction of cardiovascular disease using machine learning algorithms with relief and lasso feature selection techniques. IEEE Access, 9: 19304-19326. https://doi.org/10.1109/ACCESS.2021.3053759

[16] Mohammadi, S., Mirvaziri, H., Ghazizadeh-Ahsaee, M., Karimipour, H. (2019). Cyber intrusion detection by combined feature selection algorithm. Journal of Information Security and Applications, 44: 80–88. https://doi.org/10.1016/j.jisa.2018.11.007

[17] 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.

[18] Pham, B.T., Nguyen-Thoi, T., Ly, H.B., Nguyen, M.D., Al-Ansari, N., Tran, V.Q., Le, T.T. (2020). Extreme learning machine based prediction of soil shear strength: A sensitivity analysis using Monte Carlo simulations and feature backward elimination. Sustainability, 12(6): 2330. https://doi.org/10.3390/su12062339

[19] Bagherzadeh, F., Mehrani, M.J., Basirifard, M., Roostaei, J. (2021). Comparative study on total nitrogen prediction in wastewater treatment plant and effect of various feature selection methods on machine learning algorithms performance. Journal of Water Process Engineering, 41: 102033. https://doi.org/10.1016/j.jwpe.2021.102033

[20] Chen, C., Tsai, Y., Chang, F., Lin, W. (2020). Ensemble feature selection in medical datasets: Combining filter, wrapper, and embedded feature selection results. Expert Systems, 37(5): e12553. https://doi.org/10.1111/exsy.12553

[21] Zebari, R., Abdulazeez, A., Zeebaree, D., Saeed, J. (2020). A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction. Journal of Applied Science and Technology Trends, 1(2): 56-70. https://doi.org/10.38094/jastt1224

[22] Pes, B. (2020). Ensemble feature selection for high-dimensional data: A stability analysis across multiple domains. Neural Computing and Applications, 32(10): 5951-5973. https://doi.org/10.1007/s00521-019-04082-3

[23] Tripathi, D., Edla, D.R., Cheruku, R., Kuppili, V. (2019). A novel hybrid credit scoring model based on ensemble feature selection and multilayer ensemble classification. Computational Intelligence, 35(2): 371-394. https://doi.org/10.1111/coin.12200

[24] Bolon-Canedo, V., Alonso-Betanzos, A. (2019), Ensembles for feature selection: A review and future trends. Information Fusion, 52: 1-12. https://doi.org/10.1016/j.inffus.2018.11.008

[25] Tsai, C.F., Sung, Y.T. (2020). Ensemble feature selection in high dimension, low sample size datasets: Parallel and serial combination approaches. Knowledge-Based Systems, 203: 106097. https://doi.org/10.1016/j.knosys.2020.106097

[26] Kshirsagar, D., Kumar, S. (2021). A feature reduction based reflected and exploited DDoS attacks detection system. Journal of Ambient Intelligence and Humanized Computing, 13: 393-405. https://doi.org/10.1007/s12652-021-02907-5

[27] Alasadi, S.A., Bhaya, W.S. (2017). Review of data preprocessing techniques in data mining. Journal of Engineering and Applied Sciences, 12(16): 4102-4107.

[28] 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

[29] 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

[30] Hakkoum, H., Idri, A., Abnane, I. (2021). Assessing and comparing interpretability techniques for artificial neural networks breast cancer classification. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, 9(6): 587-599. https://doi.org/10.1080/21681163.2021.1901784