Data and Metadata. 2023; 2:173
doi: 10.56294/dm2023173

ORIGINAL

An efficient prediction system for diabetes disease based on machine learning algorithms

Un sistema eficaz de predicción de la diabetes basado en algoritmos de aprendizaje automático

Mariame Oumoulylte[1], Abdelkhalak Bahri[1], Yousef Farhaoui[2], Ahmad El Allaoui[2]

[1]Laboratory of Applied Sciences; Team: SDIC; National School of Applied Sciences Al-Hoceima, Abdelmalek Esaadi University, Tétouan, Morocco
[2]L-STI, T-IDMS, FST Errachidia, Moulay Ismail University of Meknes, Morocco

Cite as: Oumoulylte M, Bahri A, Farhaoui Y, El Allaoui A. An efficient prediction system for diabetes disease based on machine learning algorithms. Data and Metadata. 2023;2:173.https://doi.org/10.56294/dm2023173

Submitted: 22-08-2023 **Revised:** 16-10-2023 **Accepted:** 24-11-2023 **Published:** 20-12-2023

Editor: Prof. Dr. Javier González Argote

ABSTRACT

Diabetes is a persistent medical condition that arises when the pancreas loses its ability to produce insulin or when the body is unable to utilize the insulin it generates effectively. In today's world, diabetes stands as one of the most prevalent and, unfortunately, one of the deadliest diseases due to certain complications. Timely detection of diabetes plays a crucial role in facilitating its treatment and preventing the disease from advancing further. In this study, we have developed a diabetes prediction model by leveraging a variety of machine learning classification algorithms, including K-Nearest Neighbors (KNN), Naive Bayes, Support Vector Machine (SVM), Decision Tree, Random Forest, and Logistic Regression, to determine which algorithm yields the most accurate predictive outcomes. we employed the famous PIMA Indians Diabetes dataset, comprising 768 instances with nine distinct feature attributes. The primary objective of this dataset is to ascertain whether a patient has diabetes based on specific diagnostic metrics included in the collection. In the process of preparing the data for analysis, we implemented a series of preprocessing steps. The evaluation of performance metrics in this study encompassed accuracy, precision, recall, and the F1 score. The results from our experiments indicate that the K-nearest neighbors' algorithm (KNN) surpasses other algorithms in effectively differentiating between individuals with diabetes and those without in the PIMA dataset.

Keywords: Machine Learning; Healthcare; Diabetes Disease; KNN; Naive Bayes; SVM; Decision Tree; Random Forest; Logistic Regression.

RESUMEN

La diabetes es una afección médica persistente que surge cuando el páncreas pierde su capacidad de producir insulina o cuando el organismo es incapaz de utilizar eficazmente la insulina que genera. En el mundo actual, la diabetes es una de las enfermedades más prevalentes y, por desgracia, una de las más mortíferas debido a ciertas complicaciones. La detección a tiempo de la diabetes desempeña un papel crucial para facilitar su tratamiento y evitar que la enfermedad siga avanzando. En este estudio, hemos desarrollado un modelo de predicción de la diabetes aprovechando diversos algoritmos de clasificación de aprendizaje automático, como K-Nearest Neighbors (KNN), Naive Bayes, Support Vector Machine (SVM), Decision Tree, Random Forest y Logistic Regression, para determinar qué algoritmo arroja los resultados predictivos más precisos. Empleamos el famoso conjunto de datos PIMA Indians Diabetes, compuesto por 768 instancias con nueve atributos de características distintas. El objetivo principal de este conjunto de datos es determinar si un paciente padece diabetes basándose en métricas de diagnóstico específicas incluidas en la colección.

En el proceso de preparación de los datos para el análisis, implementamos una serie de pasos de preprocesamiento. La evaluación de las métricas de rendimiento en este estudio abarcó la exactitud, la precisión, la recuperación y la puntuación F1. Los resultados de nuestros experimentos indican que el algoritmo K-nearest neighbors (KNN) supera a otros algoritmos en la diferenciación efectiva entre individuos con diabetes y sin diabetes en el conjunto de datos PIMA.

Palabras clave: Aprendizaje Automático; Asistencia Sanitaria; Enfermedad de la Diabetes; KNN; Naive Bayes; SVM; Árbol de Decisión; Bosque Aleatorio; Regresión Logística.

INTRODUCTION

Cancer is an ailment distinguished by the unrestrained splitting and proliferation of cells in organs or tissues, which can lead to their spread beyond the initial location.[1] Skin cancer is a dangerous and potentially deadly type of cancer,[2,3,4] It represents the most common type that threatens human beings. In the first instance, it is visually detected, then by dermoscopic analysis, an early diagnosis makes it curable almost one hundred percent.[5,6] Precisely diagnosing skin cancer poses a significant challenge for dermatologists, even when employing dermoscopy images, due to the initial similarity in appearance among several types of skin cancer. Furthermore, even skilled dermatologists encounter limitations based on their education and experience in accurately diagnosing skin cancer. Their exposure is confined to a subset of potential skin cancer manifestations throughout their professional lifetime. Likewise, dermoscopy in the hands of less experienced dermatologists can lead to a decrease in the accuracy of skin cancer identification. Consequently, to solve the problems encountered by dermatologists there is an urgent necessity to create a swifter and more precise process for detecting and classifying skin cancer lesions.[7]

In recent years, researchers have invested substantial effort into crafting intelligent systems for applications in different fields such as: object detection,[8] emotion recognition[9] and healthcare.[10]

Deep Convolutional Neural Networks (CNNs) have demonstrated their effective-ness in performing.

MATERIALS AND METHODS

Dataset

The PIMA Indian dataset is an open-source dataset[1] that is publicly available for machine learning classification, which has been used in this work. It contains 768 patients females at least 21 years old data, and 268 of them have developed diabetes. The datasets consists of eight medical predictor variables and one target variable, Outcome which is either 1 or 0, 1 indicating that the patient is diabetic and 0, indicating that the patient is not diabetic. Figure 1 shows the ratio of people having diabetes in the PIMA Indian dataset. Table 1 demonstrates the eight features and the target of the open-source PIMA Indian dataset.

Table 1. Features of the PIMA Indian Dataset

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

Dataset Preprocessing

In the consolidated dataset, we identified some unusual zero values which indicates missing value, such as those for skin thickness, Glucose, BloodPressure, Insulin and Body Mass Index (BMI). Since these attributes cannot logically be zero, we replaced the zero values with their respective mean values. Following this adjustment, the process of counting this features becomes more straightforward. Figure 2 demonstrates the histogram of each feature. Figure 3 shows the correlation of various attributes in the PIMA dataset. The training and test dataset has been separated using the holdout validation technique, where 80 % is the training data and 20 % is the test data.[11,12,13,14]

https://doi.org/10.56294/dm2023173

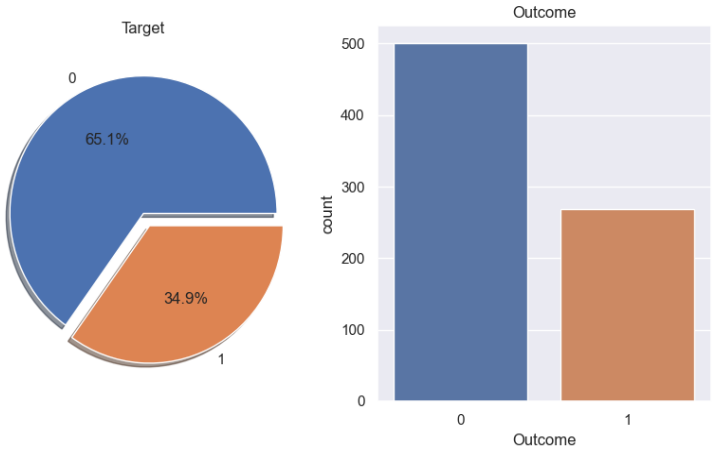

Figure 1. Percentage of people having diabetes in the PIMA Indian dataset

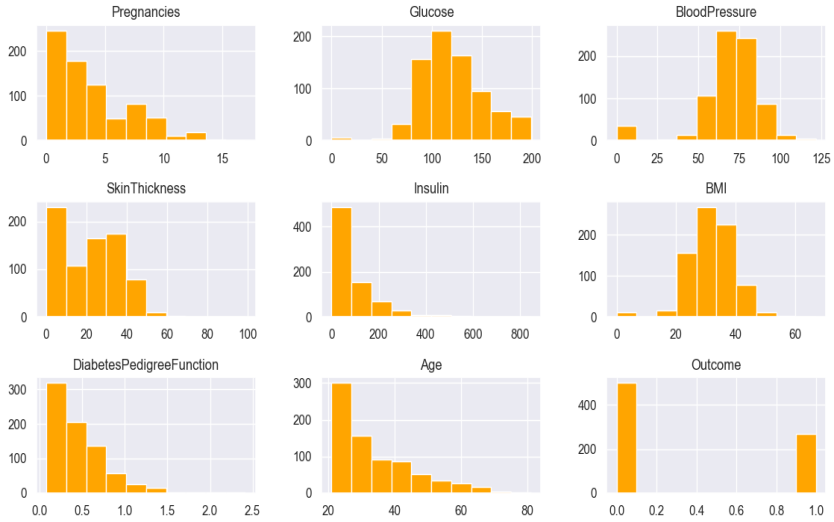

Figure 2. Histogram for each feature

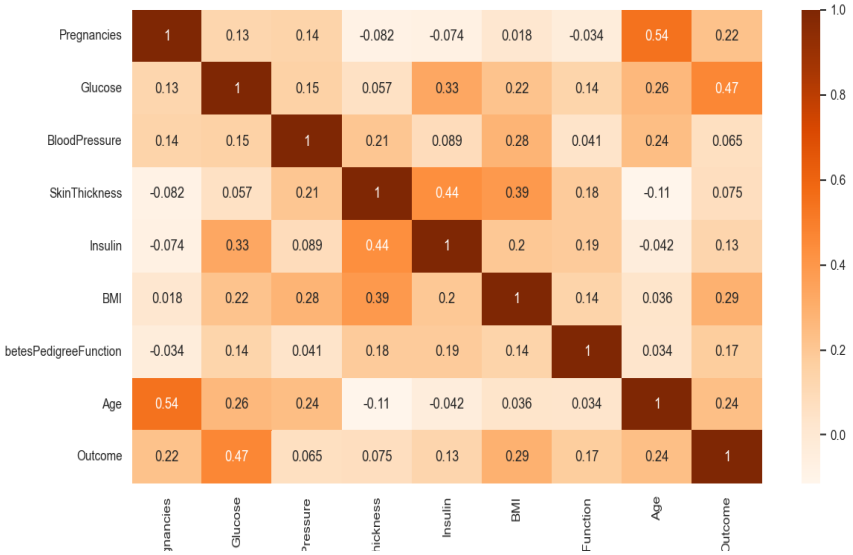

Figure 3. Correlation of Various Attributes in the PIMA Dataset

https://doi.org/10.56294/dm2023173

Table 2. Performance metrics of various classifiers

Algorithm	Class	Precision	Recall	F1-score	Accuracy	Support	Article Mujumdar A et al
KNN	0	0,83	0,91	0,87	**81 %**	107	72 %
	1	0,73	0,57	0,64		47	
Naive Bayes	0	0,82	0,87	0,84	**77 %**	107	67 %
	1	0,65	0,55	0,60		47	
SVM	0	0,82	0,91	0,86	**80 %**	107	68 %
	1	0,72	0,55	0,63		47	
Decision Tree	0	0,80	0,94	0,86	**79 %**	107	74 %
	1	0,78	0,45	0,57		47	
Random Forest	0	0,83	0,84	0,84	**77 %**	107	72 %
	1	0,63	0,62	0,62		47	
Logistic Regression	0	0,82	0,89	0,85	**79 %**	107	76 %
	1	0,68	0,55	0,61		47	

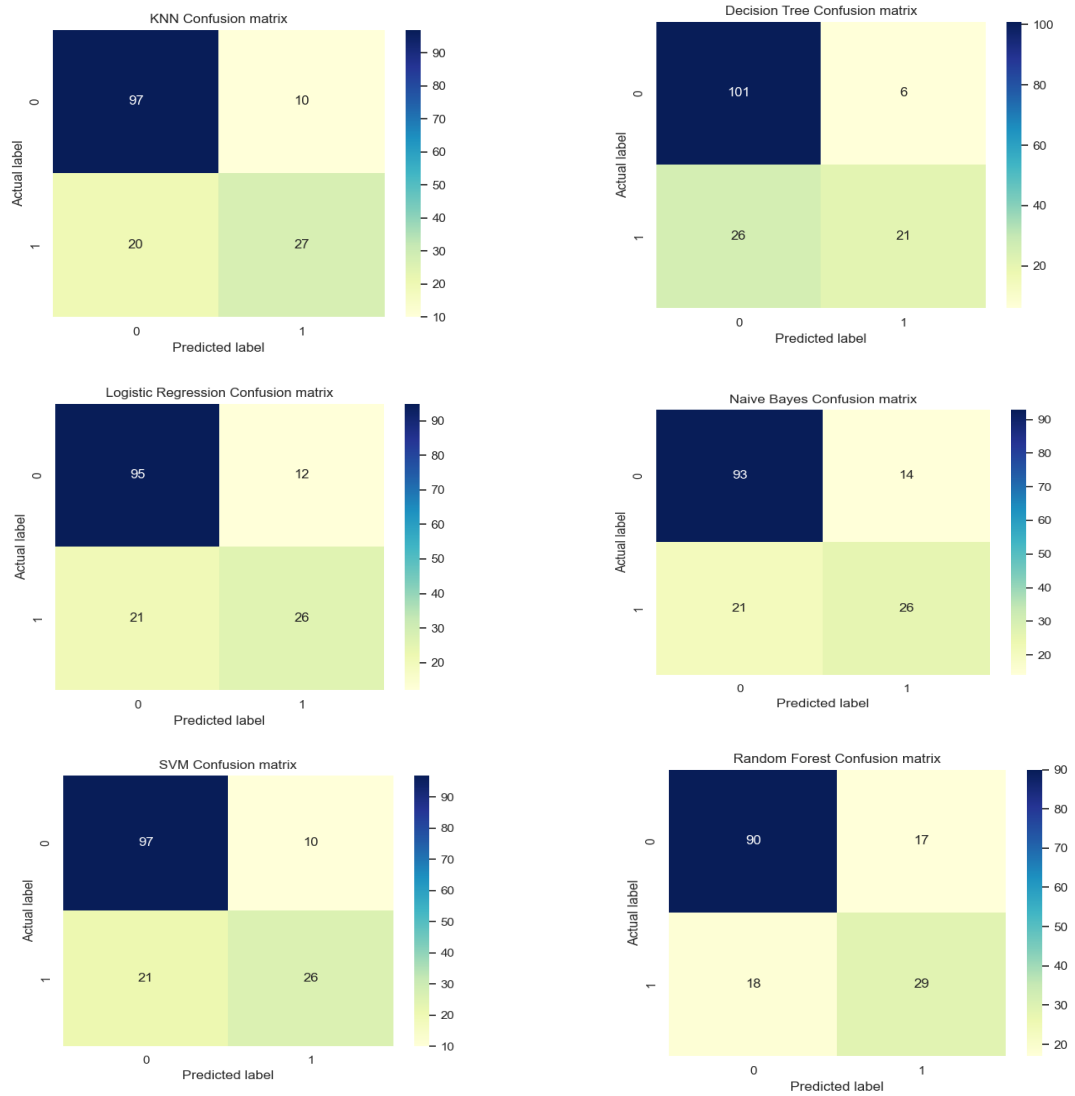

Figure 4. Confusion matrix of each model

https://doi.org/10.56294/dm2023173

Classification Algorithms

This study utilizes a range of machine learning and ensemble techniques to develop an automated diabetes prediction system, briefly outlined below. The GridSearchCV framework[2] is employed to identify optimal hyperparameter values for all machine learning models.

KNN classifier: K-Nearest Neighbors is used for both classification and regression tasks. In KNN, an object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its K nearest neighbors (K being a predefined constant). The algorithm works on the principle of proximity, where similar data points are considered to be close to each other in the feature space.[3,15,16,17]

Naive Bayes: is a probabilistic machine learning algorithm based on Bayes' theorem,[4] with the "naive" assumption of independence between features. The algorithm is named "naive" because it assumes that the presence or absence of a particular feature is independent of the presence or absence of other features, which may not always hold true in real-world data.

SVM: SVM performs supervised classification by choosing the best hyperplane.[5]

- Decision Tree: Serves as a visual representation of a learning function de-fined by a set of rules. It is employed to approximate discrete-valued target functions.[6] Gini or entropy are used to determine information gain.[7,18,19]

Random Forest: It averages the predictions of several decision trees. As a result, the random forest can be considered an ensemble learning model.[7]

Logistic Regression: The algorithm models the probability that a given in-stance belongs to a particular category. The output is transformed using the logistic function (sigmoid), which maps any real-valued number into the range between 0 and 1.[9,20,21]

Performance evaluation

We validated the models' performance by evaluating metrics such as Recall, Precision, and F1-score. see equation 1, 2, and 3:

Recall, the fraction of true positives that are correctly identified.
Precision is the fraction of retrieved instances that are relevant.
F1-score is the weighted average of Precision and Recall.

$$\text{Recall/Sensitivity} = TP/(TP+FN) \quad (1)$$

$$\text{Precision} = TP/(TP+FP) \quad (2)$$

$$\text{F1-score} = (2 * \text{Precision} * \text{Recall})/(\text{Precision} + \text{Recall}) \quad (3)$$

Where:

TP (True Positives). These instances refer to cases in which the model accurately predicted the positive class when the actual class was indeed positive;

FN (False Negatives). These occurrences correspond to situations in which the model forecasted the negative class, but the actual class was positive. In simpler terms, the model failed to identify a positive case;

FP (False Positives). These scenarios arise when the model predicts the positive class, but the actual class is negative. In such instances, the model mistakenly as-signs a negative case as positive.

RESULTS AND DISCUSSION

We used precision, recall, f1-score and classification accuracy to evaluate various ML models as detailed in Table 2. The classification model performance is visually depicted in Figure 4 through the confusion matrix.

Table 2 compares different performance metrics of various classifiers for the merged dataset. As per the provided table, the KNN classifier demonstrated superior overall performance. It achieved an 81 % accuracy, along with precision, recall, and F1-score of 0,83, 0,91, and 0,87, respectively, for the category of non-diabetics. However, its performance was comparatively lower for the category of diabetics, with precision, recall, and F1-score values of 0,73, 0,57, and 0,64, respectively.

Figure 4 depicts the confusion matrix for various ML models. According to this figure KNN technique correctly classified 124 instances with TP = 97 and TN = 27.[22]

CONCLUSION

Diabetes, known to impact life expectancy and overall quality of life, can be mitigated by early prediction, minimizing risks and long-term complications. This work introduces an automatic diabetes prediction system employing diverse machine learning approaches. Utilizing the open-source PIMA Indian dataset, this study aims to contribute to the early identification and management of this chronic disorder. This research reported different performance metrics, that is, precision, recall, accuracy, and F1 score, for various machine learning models. The KNN classifier achieved the best performance with 81 % accuracy.

REFERENCES

1. PIMA Indians Diabetes Database. (2016, October 6). https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

2. sklearn.model_selection.GridSearchCV. (n.d.). Scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

3. What is the k-nearest neighbors algorithm? | IBM. (n.d.). https://www.ibm.com/topics/knn

4. Ray, S. (2023, December 1). Naive Bayes Classifier explained: Applications and practice problems of Naive Bayes Classifier. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/

5. Support Vector Machine (SVM) explained. (n.d.). MATLAB & Simulink. https://se.mathworks.com/discovery/support-vector-machine.htmlhttps://www.ibm.com/topics/decision-trees

6. Aurélien, G.: Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly Media, Inc., Sebastopol, CA

7. Towards data science. Available https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148

8. Beghriche, T., Djerioui, M., Brik, Y., Attallah, B., & Belhaouari, S. B. (2021). An efficient prediction system for diabetes disease based on deep neural network. Complexity, 2021, 1–14. https://doi.org/10.1155/2021/6053824

9. Tasin, I., Nabil, T. U., Islam, S., & Khan, R. (2022). Diabetes prediction using machine learning and explainable AI techniques. Healthcare Technology Letters, 10(1–2), 1–10. https://doi.org/10.1049/htl2.12039

10. AD Association. Classification and diagnosis of diabetes: standards of medical care in diabe-tes-2020. Diabetes Care. 2019. https://doi.org/10.2337/dc20-S002.

11. Sanal MG, Paul K, Kumar S, Ganguly NK. Artificial intelligence and deep learning: the future of medicine and medical practice. J Assoc Physicians India. 2019;67(4):71–3.

12. Muhammad LJ, Algehyne EA, Usman SS. Predictive supervised machine learning models for diabetes mellitus. SN Comput Sci. 2020;1(5):1–10. https://doi.org/10.1007/s42979-020-00250-8.

13. Hasan, M.K., Alam, M.A., Das, D., Hossain, E., Hasan, M.: Diabetes prediction using en-sembling of different machine learning classifiers. IEEE Access 8, 76516–76531, (2020)

14. Pranto, B., et al.: Evaluating machine learning methods for predicting diabetes among female patients in Bangladesh. Information 11, 1–20 (2020)

15. Jackins, V., Vimal, S., Kaliappan, M., Lee, M.Y.: AI-based smart prediction of clinical dis-ease using random forest classifier and Naive Bayes. J. Supercomput. 77, 5198–5219 (2021)

16. F. Mohanty, S. Rup, and B. Dash, "Automated diagnosis of breast cancer using parameter optimized kernel extreme learning machine," Biomedical Signal Processing and Control, vol. 62, pp. 102–108, 2020.

17. E. Martinez-R´ıos, L. Montesinos, M. Alfaro-Ponce, and L. Pecchia, "A review of machine learning in hypertension detection and blood pressure estimation based on clinical and phys-iological data," Biomedical Signal Processing and Control, vol. 68, Article ID 102813, 2021.

18. H. Naz and S. Ahuja, "Deep learning approach for diabetes prediction using PIMA Indian dataset," Journal of Diabetes & Metabolic Disorders, vol.19(1), pp.391-403, 2020.

19. F. Mohanty, S. Rup, and B. Dash, "Automated diagnosis of breast cancer using parameter optimized kernel extreme learning machine," Biomedical Signal Processing and Control, vol. 62, pp. 102–108, 2020.

20. Oumoulylte, M., El Allaoui, A., Farhaoui, Y., Amounas, F. & Qaraai, Y. Deep Learning Algorithms for Skin

Cancer Classification. Artificial Intelligence and Smart Environment. ICAISE 2022. Lecture Notes in Networks and Systems, Springer, Cham, 2022, vol. 635, pp. 345-351. DOI: 10.1007/978-3-031-26254-8_49.

21. A Novel Diabetes Healthcare Disease Prediction Framework Using Machine Learning Techniques. Krishnamoorthi R, Joshi S, Almarzouki HZ, Shukla PK, Rizwan A, Kalpana C, Tiwari B.J Healthc Eng. 2022 Jan 11;2022:1684017. doi: 10.1155/2022/1684017. eCollec-tion 2022.

22. Mujumdar, A., & Vaidehi, V. (2019). Diabetes Prediction using Machine Learning Algo-rithms. Procedia Computer Science, 165, 292–299. https://doi.org/10.1016/j.procs.2020.01.047

FINANCING

No financing.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

AUTHORSHIP CONTRIBUTION

Conceptualization: Mariame Oumoulylte, Abdelkhalak Bahri, Yousef Farhaoui, Ahmad El Allaoui.
Research: Mariame Oumoulylte, Abdelkhalak Bahri, Yousef Farhaoui, Ahmad El Allaoui.
Drafting - original draft: Mariame Oumoulylte, Abdelkhalak Bahri, Yousef Farhaoui, Ahmad El Allaoui.
Writing - proofreading and editing: Mariame Oumoulylte, Abdelkhalak Bahri, Yousef Farhaoui, Ahmad El Allaoui.