African Journal of Advanced Pure and Applied Sciences (AJAPAS)
Online ISSN: 2957-644X
Volume 3, Issue 3, 2024, Page No: 373-377
Website: *https://aaasjournals.com/index.php/ajapas/index*
معامل التأثير العربي 2023: (1.55) SJIFactor 2023: 5.689 ISI 2022-2023: 0.557
Special issue: First Libyan Conference on Technology and Innovation (LCTI-2024), Benghazi, Libya

دراسة تأثير اختيار الميزات على التنبؤ بمرض السكري

محمد شنتال[3،1]*، المهدي الشريف[3،2]، عمر أحميد[1]
[1] قسم الحاسوب، كلية العلوم التقنية، سبها، ليبيا
[2] قسم الحاسوب، كلية تقنية المعلومات، جامعة سبها، سبها، ليبيا
[3] مركز التطوير المعلوماتي، جامعة سبها، سبها، ليبيا

The Impact Of Feature Selection On Diabetes Prediction

Mohammed Shantal [1,3] *, Almahdi Alshareef [1,2], Omar Ahmid [1]
[1] Computer Science Department, College of Technology Science, Sebha, Libya
[2] Computer Science Department, Sebha University, Sebha, Libya
[3] Information Development Center, Sebha University - Sebha, Libya

Corresponding author: moh.shantal@sebhau.edu.ly

Received: March 01, 2024	Accepted: May 05, 2024	Published: May 30, 2024

الملخص

السكري هو حالة استقلابية مزمنة تتميز بمستويات غلوكوز الدم غير الطبيعية نتيجة لاستخدام الأنسولين غير الفعال أو الإنتاج غير الكافي، دفع العديد من الجهود الأكاديمية إلى تصميم نماذج تنبؤية موثوقة باستخدام خوارزميات التعلم الآلي. إزالة الميزات الزائدة من مجموعات البيانات الضخمة ضرورية لتحسين كفاءة النماذج التنبؤية المدعومة بالبيانات. يهدف هذا العمل إلى دراسة تأثير طرق اختيار الميزات على تدريب ودقة مصنفات البيانات. تمت مقارنة ثلاث طرق لاختيار الميزات وهي f_classif وchi^2 وRFE مع مجموعة البيانات الكاملة. بالإضافة إلى ذلك، تم استخدام تسع مصنفات (Naïve Bayes, *k*-NN, Decision Tree, Random Forest, Support Vector Machine, Gradient Boosting، ExtraTrees, AdaBoost, Multilayer Perceptron) لتقييم دقة كل طريقة لاختيار الميزات. ومن النتائج، حصلت RFE على أفضل دقة بين استراتيجيات اختيار الميزات الأخرى، حيث حقق معظم المصنفات أفضل نتائجها باستخدام RFE، مع حصول 5 من 9 مصنفات على أفضل نتائجها باستخدام RFE.

الكلمات المفتاحية: اختيار الميزات، التنبؤ بمرض السكري، Filter selection، wrapper selection، f_classif، chi^2، RFE

Abstract

Diabetes, a chronic metabolic condition characterized by abnormal blood glucose levels due to either ineffective insulin utilization or inadequate production, has prompted numerous academic efforts to devise dependable prediction models using machine learning (ML) algorithms. Removing redundant features from massive datasets is of paramount importance in improving the efficiency of data-driven predictive models. This work aims to study the impact of feature selection (FS) methods on classifier training and accuracy. Three FS methods, f_classif, chi^2, and RFE, were compared with the full dataset. Additionally, nine classifiers (Naïve Bayes, *k*-NN, Decision Tree, Random Forest, Support Vector Machine, Gradient Boosting, Multilayer perceptron, AdaBoost, and ExtraTrees) were employed to evaluate the accuracy of each FS method. From the results, RFE obtained the best accuracy across other FS strategies, with most classifiers achieving their best results using RFE, with 5 out of 9 classifiers obtaining their best results using RFE.

Keywords: Feature Selection, Diabetes prediction, Filter selection, wrapper selection, f_classif, chi^2, RFE

Introduction

One of the common prediction models is early prediction of diabetics which has a high impact on the health systems and societies [1]. Over time, many researchers have aimed to create accurate diabetes prediction models, but the field faces persistent challenges due to insufficient datasets and prediction methods. As a result, scholars are increasingly utilizing big data analytics and ML approaches to tackle these obstacles [2]. Sisodia and Sisodia [3] developed a diabetes prediction system, employing three machine learning algorithms Naïve Bayes (NB), Support Vector Machine (SVM), and Decision Tree (DT), with particularly effective results observed in diabetes prediction using the NB algorithm. The work's aim of Zhu, et al. [4] was to develop a robust diabetes prediction model. They introduced a novel approach integrating PCA for dimensionality reduction, k-means for clustering, and logistic regression for classification, alongside preprocessing steps applied to the dataset after analysis the data. The experimental results demonstrated enhanced accuracy following these modifications.

Hasan, et al. [5] focused on diabetes prediction using an ensemble model developed from the PIMA dataset, with preprocessing enhancing dataset quality by addressing outliers and missing values. Attribute selection methods, such as correlation-based selection, improve attribute-target outcome correlation. The proposed framework outperforms others in AUC, with the combination of boosting classifiers (AB and XB) showing promise for diabetes prediction, particularly when coupled with the proposed preprocessing techniques.

In the other side, FS constitutes a fundamental preprocessing step in machine learning endeavors, proven effective for reducing dimensionality and eliminating irrelevant or redundant features [6]. FS has been divided into three categories, Filter, Wrapper, and Embedded methods [7]. Filter methods assess feature importance independently, disregarding their interactions in predictive models, typically by computing scores based on statistical measures like correlation or mutual information with the target feature. Examples include Relief, Mutual information (MI), Chi-squared (chi^2), and Correlation-based features [8]. Wrapper methods assess feature subsets by iteratively training models, and selecting the best-performing subset. Random Forest (RF) is a prime example, employing multiple DTs on varied feature subsets, with the most important features determined by consensus across trees [9]. The final category is Embedded FS methods which are integrated directly into the learning algorithm, rather than being a separate step, thus seamlessly incorporating selection into the modeling process. Examples include DT algorithms like CART, C4.5, and RF [10]. See Figure 1.

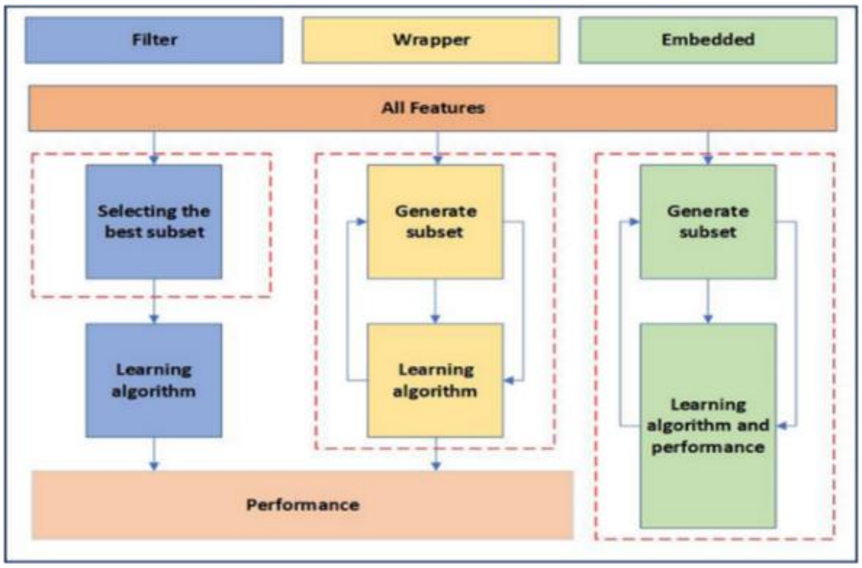

Figure 1: The types of FS methods.

In this work, we aim to study and analyze the impact of feature selection (FS) on diabetes prediction, where various FS methods from the filter and wrapper categories are applied to select important features from the Diabetes dataset. Next, the results will be evaluated to determine if the FS strategies improve the accuracy of machine learning classification methods.

Material and methods

The proposed methodology is illustrated in Figure 2 below, depicting the research flow involved in constructing the model. Initially, multiple datasets including the Full dataset, as well as three derived datasets from FS methods chi^2, f_classifier, RFE), are created. Subsequently, each dataset is divided into a training set and a test set for model training and evaluation. Five classifiers are employed to predict classes after model training, followed by accuracy assessment for each classifier. The influence of FS is examined by comparing the performance between using the full dataset and selected feature sets to discern any potential enhancements.

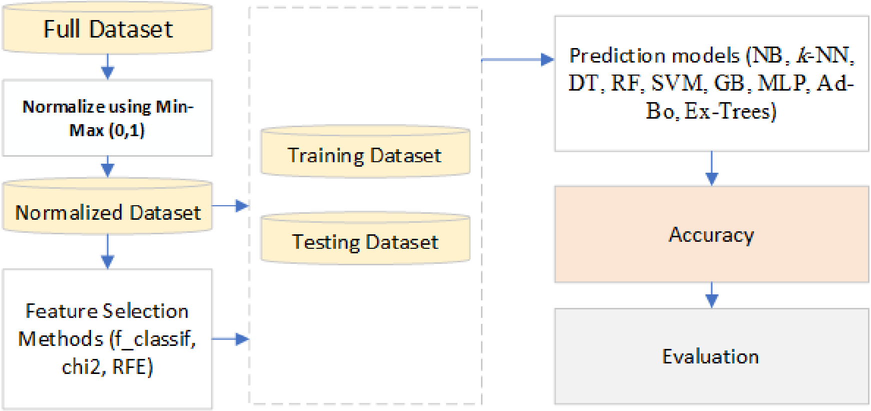
Full Dataset
Normalize using Min-Max (0,1)
Normalized Dataset
Feature Selection Methods (f_classif, chi2, RFE)
Training Dataset
Testing Dataset
Prediction models (NB, k-NN, DT, RF, SVM, GB, MLP, Ad-Bo, Ex-Trees)
Accuracy
Evaluation

Figure 2: The general methodology of the study.

Feature Selection methods

In this work, three FS methods have been chosen to employ on the dataset. First is the f_classif (FC), primarily for classification, which calculates ANOVA F-values for each feature relative to the target variable, selecting those with the highest values to signify significant differences across classes, under the assumption of normal distribution and equal class variance [11]. Second is the chi^2 method, common in classification tasks with categorical features, assesses the statistical significance of the association between each feature and the target variable using the chi^2 statistic [12]. Finally, Recursive Feature Elimination (RFE) is the third method that has been used in this study. RFE is a FS technique that recursively removes features from a model based on their importance, iteratively refining the model until the optimal subset of features is determined [13]. RF classifier is the model that has been used with RFE method. Also, the five best features have been selected in each FS method.

Datasets

This study utilized the PIMA Indians Diabetes dataset, comprising 768 female diabetic patients from the Pima Indian community. The dataset includes 268 diabetic patients (considered positive) and 500 non-diabetic patients (considered negative), with eight distinct attributes [14, 15].

Classifiers method

This study employs nine classifiers: NB, k nearest neighbors *k*-NN, DT, RF, SVM, Gradient Boosting (GB), ,Multilayer perceptron (MLP), AdaBoost (Ad-Bo), and ExtraTrees (Ex-Trees) [16]. NB utilizes Bayes theorem with an independence assumption for straightforward probabilistic prediction [17]. The *k*-NN algorithm is based on selecting the nearest k neighbours, among which votes are cast to determine the most frequently occurring class [18]. DT is a hierarchical structures used for decision-making, where each internal node represents a decision based on a feature, and each leaf node represents a class label [19]. RF utilizes multiple DT to enhance predictive accuracy [20]. SVM is a powerful supervised learning algorithm used for classification and regression tasks, which works by finding the optimal hyperplane that separates data points into different classes [21]. MLP classifier utilizes multiple layers and non-linear activation functions for discerning non-linearly separable data. SVM is employed for classification, regression, and FS, maximizing margin to enhance generalization [22]. ExtraTrees reduces computational resource usage, employing the entire training set and optimal attribute selection to prevent overfitting and improve performance [16].

Evaluation strategy

In this section, we present standard metrics used to evaluate the effectiveness of ML algorithms in classification and prediction endeavours. Among these metrics is the accuracy, which compares the predicted labels generated by the classifier with the actual labels obtained from the dataset.

Results and discussion

As seen in Table 1, it's evident that the performance of classifiers varies across different FS methods. For example, the NB classifier achieved its highest accuracy with RFE, while its lowest accuracy was observed with chi^2. Similarly, the *k*-NN classifier performed best with FC, while chi^2 resulted in the lowest accuracy. For DT, SVM, RF, and Ad-Bo classifiers, RFE yielded the highest accuracy, while chi^2 resulted in the lowest. However, chi^2 provided the best accuracy with the GB classifier, and the full dataset yielded the highest accuracy with the Ex-Trees classifier. Overall, there are slight improvements in accuracy when using certain FS methods, such as FC and RFE, compared to others. This suggests that the choice of FS method can modestly impact classifier performance, with certain methods potentially offering slightly better accuracy for specific classifiers.

Table 1 Result of Classifiers accuracy using full dataset vs four FS strategies. (**Bold** values present the best accuracy while the underlined values present the lowest accuracy).

Classifier	FD Methods			
	FD	FC	Chi^2	RFE
NB	75.52%	76.12%	73.79%	**76.77%**
k-NN	73.91%	**75.10%**	73.61%	73.95%
DT	73.93%	74.13%	72.11%	**74.19%**
RF	76.47%	76.28%	74.78%	**76.60%**
SVM	77.00%	76.92%	75.95%	**77.43%**
GB	75.05%	**75.48%**	74.76%	75.09%
MLP	77.03%	77.03%	**77.37%**	77.19%
Ad-Bo	75.70%	76.07%	74.88%	**76.11%**
Ex-Trees	**76.46%**	75.85%	74.91%	76.11%

Conclusion

In summary, our analysis reveals variations in classifier performance across different FS methods. While certain classifiers, like NB and *k*-NN, showed notable differences in accuracy depending on the method employed, others, such as SVM and Ad-Bo, consistently performed best with RFE. Surprisingly, chi^2 yielded the highest accuracy for the GB classifier, suggesting method-specific nuances. Our findings emphasize the importance of method selection in optimizing classifier performance, with slight improvements observed when using certain FS techniques.

References

[1] H. Zhou, R. Myrzashova, and R. Zheng, "Diabetes prediction model based on an enhanced deep neural network," *EURASIP Journal on Wireless Communications and Networking*, vol. 2020, no. 1, p. 148, 2020/07/17 2020, doi: 10.1186/s13638-020-01765-7.

[2] H. B. Kibria, M. Nahiduzzaman, M. O. F. Goni, M. Ahsan, and J. Haider, "An Ensemble Approach for the Prediction of Diabetes Mellitus Using a Soft Voting Classifier with an Explainable AI," *Sensors*, vol. 22, no. 19, p. 7268, 2022. [Online]. Available: https://www.mdpi.com/1424-8220/22/19/7268.

[3] D. Sisodia and D. S. Sisodia, "Prediction of Diabetes using Classification Algorithms," *Procedia Computer Science*, vol. 132, pp. 1578-1585, 2018/01/01/ 2018, doi: https://doi.org/10.1016/j.procs.2018.05.122.

[4] C. Zhu, C. U. Idemudia, and W. Feng, "Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques," *Informatics in Medicine Unlocked*, vol. 17, p. 100179, 2019/01/01/ 2019, doi: https://doi.org/10.1016/j.imu.2019.100179.

[5] M. K. Hasan, M. A. Alam, D. Das, E. Hossain, and M. Hasan, "Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers," *IEEE Access*, vol. 8, pp. 76516-76531, 2020, doi: 10.1109/ACCESS.2020.2989857.

[6] E. C. Blessie and E. Karthikeyan, "Sigmis: A Feature Selection Algorithm Using Correlation Based Method," *Journal of Algorithms & Computational Technology*, vol. 6, no. 3, pp. 385-394, 2012, doi: 10.1260/1748-3018.6.3.385.

[7] Z. M. Hira and D. F. Gillies, "A review of feature selection and feature extraction methods applied on microarray data," *Advances in bioinformatics*, vol. 2015, 2015.

[8] S.-i. Kim, Y. Noh, Y.-J. Kang, S. Park, J.-W. Lee, and S.-W. Chin, "Hybrid data-scaling method for fault classification of compressors," *Measurement*, vol. 201, p. 111619, 2022/09/30/ 2022, doi: https://doi.org/10.1016/j.measurement.2022.111619.

[9] A. Bommert, X. Sun, B. Bischl, J. Rahnenführer, and M. Lang, "Benchmark for filter methods for feature selection in high-dimensional classification data," *Computational Statistics & Data Analysis*, vol. 143, p. 106839, 2020/03/01/ 2020, doi: https://doi.org/10.1016/j.csda.2019.106839.

[10] A. Jović, K. Brkić, and N. Bogunović, "A review of feature selection methods with applications," in *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 25-29 May 2015 2015, pp. 1200-1205, doi: 10.1109/MIPRO.2015.7160458.

[11] C. Vong, T. Theptit, V. Watcharakonpipat, P. Chanchotisatien, and S. Laitrakun, "Comparison of feature selection and classification for human activity and fall recognition using smartphone sensors," in *2021 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunication Engineering*, 2021: IEEE, pp. 170-173.

[12] S. K. Dey, K. M. M. Uddin, H. M. H. Babu, M. M. Rahman, A. Howlader, and K. A. Uddin, "Chi2-MI: A hybrid feature selection based machine learning approach in diagnosis of chronic kidney disease," *Intelligent Systems with Applications,* vol. 16, p. 200144, 2022.

[13] R. S. Arslan, "Comparison of Feature Selection Methods in Security Analysis of Android," in *2021 6th International Conference on Computer Science and Engineering (UBMK)*, 15-17 Sept. 2021 2021, pp. 1-5, doi: 10.1109/UBMK52708.2021.9558984.

[14] V. Chang, J. Bailey, Q. A. Xu, and Z. Sun, "Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms," *Neural Computing and Applications,* vol. 35, no. 22, pp. 16157-16173, 2023/08/01 2023, doi: 10.1007/s00521-022-07049-z.

[15] P. V. Sankar Ganesh and P. Sripriya, "A Comparative Review of Prediction Methods for Pima Indians Diabetes Dataset," Cham, 2020: Springer International Publishing, in Computational Vision and Bio-Inspired Computing, pp. 735-750.

[16] M. Alabadla, F. Sidi, I. Ishak, H. Ibrahim, L. S. Affendey, and H. Hamdan, "ExtraImpute: A Novel Machine Learning Method for Missing Data Imputation," *journal of advances in information technology,* vol. 13, no. 5, pp. 470-476, OCT 2022, doi: 10.12720/jait.13.5.470-476.

[17] N. Salmi and Z. Rustam, "Naïve Bayes Classifier Models for Predicting the Colon Cancer," *IOP Conference Series: Materials Science and Engineering,* vol. 546, no. 5, p. 052068, 2019/06/01 2019, doi: 10.1088/1757-899x/546/5/052068.

[18] N. Hasdyna, B. Sianipar, and E. M. Zamzami, "Improving The Performance of K-Nearest Neighbor Algorithm by Reducing The Attributes of Dataset Using Gain Ratio," *Journal of Physics: Conference Series,* vol. 1566, p. 012090, 2020/06 2020, doi: 10.1088/1742-6596/1566/1/012090.

[19] S. Nikfalazar, C.-H. Yeh, S. Bedingfield, and H. A. Khorshidi, "Missing data imputation using decision trees and fuzzy clustering with iterative learning," *Knowledge and Information Systems,* vol. 62, no. 6, pp. 2419-2437, 2020.

[20] M. M. A. Shibly, T. A. Tisha, and M. M. I. Mazumder, "Predicting early readmission of diabetic patients: Toward interpretable models," in *International Conference on Communication, Computing and Electronics Systems: Proceedings of ICCCES 2020*, 2021: Springer, pp. 185-200.

[21] M. J. Bazrkar and S. Hosseini, "Predict Stock Prices Using Supervised Learning Algorithms and Particle Swarm Optimization Algorithm," *Computational Economics,* vol. 62, no. 1, pp. 165-186, 2023/06/01 2023, doi: 10.1007/s10614-022-10273-3.

[22] S. Prasad, T. S. Savithri, and I. V. M. Krishna, "Comparison of Accuracy Measures for RS Image Classification using SVM and ANN Classifiers," *International Journal of Electrical and Computer Engineering,* vol. 7, no. 3, p. 1180, 2017.