International Journal of Engineering, Science and Humanities
An international peer reviewed, refereed, open-access journal
Impact Factor: 6.5 www.ijesh.com ISSN: 2250 3552

Comparative Study of Machine Learning Pre-processing Techniques for Diabetes Mellitus Prediction

Jay Pandit

Research Scholar, Maharaja's College, Ernakulam (Kochi)

Abstract:
Diabetes Mellitus is a chronic and life-threatening disease that requires timely and accurate diagnosis to prevent complications. Machine Learning (ML) techniques have become critical in healthcare for early disease prediction, but their performance heavily depends on the quality of input data. Data pre-processing, including cleaning, feature selection and handling missing values, significantly enhances predictive accuracy. This paper provides a comparative analysis of various pre-processing methods applied to diabetes prediction datasets, including the Pima Indian Diabetes dataset. Techniques such as Average Weighted Objective Distance (AWOD), feature engineering pipelines, wrapper-based feature selection, Support Vector Machines, Random Forests, fuzzy SVM, ANFIS-based imputations, clustering and hybrid approaches are reviewed. Results from previous studies show that optimized feature selection and tailored pre-processing pipelines can boost prediction accuracy to over 98% in some cases. The paper concludes that while no single method is universally applicable, pre-processing plays a crucial role in improving the robustness and accuracy of diabetes prediction models. Future work should focus on adaptive hybrid models combining multiple techniques to handle heterogeneous healthcare datasets.

Keywords: Diabetes Mellitus, Machine Learning, Data Pre-processing, Feature Selection, Prediction Accuracy, Healthcare Analytics

Introduction:

A continuous sickness or illness in humans is referred to as a chronic disorder or remains its effect for a long time. Diabetes, called Diabetes Mellitus, affects human beings at an early stage of life and this is a kind of chronic disease, which impacts for a long duration or entire life. When someone has obesity, their level of glucose in their blood is higher because their body either produces too little hormone or their cells do not react to it correctly. It impacts the quality of life, life expectancy, or other chronic diseases like obesity, hypertension, angina, etc. Cardiovascular risks are the well-known adverse impact of diabetes. As per WHO, around 422 Every year, 1.5 million individuals worldwide lose their lives to diabetes, affecting a million people worldwide. Pre-processing data can improve the prediction model's accuracy, but it's a complicated procedure. Pre-processing models that already exist can be enhanced, particularly for intricate datasets. Diabetes of both types are the two main forms. Type 1 diabetes occurs when the body's inability to create insulin is caused

International Journal of Engineering, Science and Humanities
An international peer reviewed, refereed, open-access journal
Impact Factor: 6.5 www.ijesh.com ISSN: 2250 3552

by malfunctioning body cells. A person's lifestyle plays a role in the development of type 2 diabetes. However, be it Type 1 or Type 2, both are severe in nature and high glucose in the body can cause several other diseases. There is another type of temporary diabetes, specific to women during pregnancy. When insulin-blocking hormones develop in the body, this is called Gestational Diabetes. Early diagnosis of Diabetes is important to avoid serious and life-threatening complications in patients. Disease detection from symptoms or prediction based on certain parameters is quite a complex procedure as many major symptoms might indicate some common disease but there might be chances minor symptoms may indicate some altogether different disease. Different studies have been done for specific diseases and machine learning algorithms have been implemented to analyze different symptoms and parameters. Many models have been implemented to predict Diabetes. Accuracy in predicting Diabetes is the major key for any model. Pre-Processing of the dataset can increase the accuracy of the result, especially when the dataset is complex and has redundant and missing values.

Literature Review

In the current era, Machine learning algorithms are helping to take decisions in many domains. It's critical that patients with diabetes receive an early diagnosis in order to prevent major, sometimes fatal consequences. Identifying a disease from its symptoms or making a forecast based on specific criteria. Many algorithms have been implemented to improve the accuracy of decisions. However, still, there are lots of areas where these decision-making algorithms can be improved. Disease detection is the domain, where many researchers have worked in the last few years, still, there are many things, which are uncovered and further work is required. Specific to Diabetes, many algorithms have been implemented by many researchers using data mining and machine learning, the accuracy rate of this algorithm is good enough, but still, there are many chances of error while predicting the diseases.

[1] suggested a two-step process that considered each patient's unique health characteristics for predicting type 2 diabetes using an average-based weighted objective distance (AWOD) model. I. AWOD Assessment.

II. Assessment of the Predictions Model Based on AWOD.

As per mentioned in this study, AWOD is dependent on important and unimportant elements that have a genuine impact on the forecast. The underlying principle of this approach is the individual's health conditions considered by the health care professional while diagnosis. Pima Indian Diabetes and Mendeley Data for Diabetes datasets have been used to evaluate the performance of the model. [2] has used the The UCI Repository's diabetes dataset. This dataset is gender-neutral. Using the Crow Search algorithm, 19 features are selected. To achieve higher accuracy, they design various feature engineering pipelines and achieved 98.46% accuracy for 19 features. When using a

International Journal of Engineering, Science and Humanities
An international peer reviewed, refereed, open-access journal
Impact Factor: 6.5 www.ijesh.com ISSN: 2250 3552

construction pipeline consisting of SVD (20) + CS and SVD (13) + Filter (8), respectively, same accuracy is obtained with feature set sizes of 13 and 8 features.

[3] presented a model to predict early diabetes patients using machine learning-based algorithms. This model is using a wrapper-based feature selection to optimize the Multilayer Perceptron algorithm to reduce the input dataset attributes. it can be concluded that by reducing the features, higher accuracy can be achieved.

[4] in "" categorization models for likelihood prediction of diabetes at an early stage using feature selection" demonstrated how adding variables to a model improves it. Also, feature selection removed the redundant data from the dataset. This methodology was implemented as Waikato Environment of Knowledge Analysis (WEKA).

This model has been evaluated based on F Measure, Precision-Recall curve and Receiver Operating Characteristics Area under the curve and has shown better results. This study has also verified that feature selection and filtering of redundant attributes are required to increase the accuracy level of any model.

[5] verified the outcome of multiple conventional algorithms. [5] has found that the accuracy of the SVM approach is 93%, however, Random Forest Approach accuracy was 77%. The accuracy of any approach is mostly dependent on the dataset. Following is the result for each algorithm. [6] has designed a prediction algorithm in "Analysis of diabetes mellitus for early prediction using optimal features selection" and the results are close to clinical outcomes. In this approach, attributes for the dataset are arranged in a sequence and based on the Co-relation value. Accuracy has been increased with conventional machine learning algorithms with a new sequence of attributes in an optimal dataset. The decision tree algorithm and random forest have given the best accuracy results which are 98.2% and 98% respectively.

[7] utilized Fuzzy Support Vector Machine and F-Score selection of features to identify and categorize diabetes mellitus and then used this technique on the Pima Indian Diabetes dataset. In this model, pre-processing has been done for the dataset and removed the attributes which are having maximum missing values and Set an upper limit of less than or equal to 5% for each feature's tolerances percentages. Removed two attributes from provided eight attributes in the dataset.

The second stage involved measuring the component that distinguishes two classes using an easy method i.e., F-Score and removed the redundant and irrelevant features. After cleaning all noisy and redundant data in pre-processing and feature selection, applied the Fuzzy SVM techniques and attained 89.02% accuracy in results. Additionally, this method maintains a suitable level of accuracy while offering an optimum count of fuzzy rules.

International Journal of Engineering, Science and Humanities
An international peer reviewed, refereed, open-access journal
Impact Factor: 6.5 www.ijesh.com ISSN: 2250 3552

[8] proposed three states algorithm to predict diabetes. In the first stage, pre-processed the dataset and filled in the missing values using the ANFIS model and proposed the Enhanced Inertia Weight Binary preparation, removed the noisy and inconsistent data and then used the K-Nearest Neighbour algorithm to divide information into k groups. Assign the relevant attributes to each cluster according to the center of it.

In the second classification stage, a Decision Tree based classification approach is used to assign the appropriate class to each dataset. The Pima Indians Diabetes dataset achieved a 98.7% classification result as compared to existing conventional machine learning algorithms.

[10] had pre-processed the PMI dataset to fill the missing values and feature selection. To fill the missing values using the mean, median and K-Nearest Neighbour values. For feature selection, used a combination of attribute subset selection techniques along provide an effective preprocessing technique for better accuracy. Finally used the SVM classification algorithm to classify diabetes.

With several experiments, they found that the combination of replacement with mean as missing value analysis method and optimized selection using genetic algorithms as attribute subset selection method outperformed compared to other methods.

This algorithm is based on the principle that all the features don't contribute to predicting the effective results, even filtering redundant and unnecessary attributes can increase the accuracy. From this study,

Bat Algorithm (EIWBBA) with K-Means clustering. This method has been intended to encourage the hearing behavior of bats, which causes them to chase after prey by increasing the frequency and decreasing the loudness of their ultrasonic signal.

Feature selection was carried out using the Improved Distributed Kernel-based Principal Component Analysis (IDKPCA). Pre-processing and feature selection for this work were conducted using the Pima Indian Diabetic (PID) dataset.

In the third step, applied this pre-processed and filtered dataset to the Support Vector Machine (SVM) classification algorithm for classification and achieved 91.87% accuracy in the algorithm.

A two-stage method to predict diabetes was proposed by [9] in "An Accurate Diabetes Prediction System Based on K-means Clustering and Proposed Classification Approaches." In the first stages of data

"Preparation compensatory methods for enhanced categorization of medical datasets with imbalances" [11] has studied and identified the factors which affect the classification results negatively and proposed a method to balance the uneven and complex class distribution datasets. Three binary and two multiclass datasets have been selected for several data pre-processing methods. There is a combination of classification and certain pre-processing techniques, which

outperforms other techniques and there will be a reduction of correctly predicted labels in datasets having complex distribution and large features.

"Impact of preprocessing on medical data classification", [12] showed that pre-processing of the dataset with the right combination has a significant impact on the classification of a dataset and included discretization of numeric values, attribute subsets and missing values in pre-processing operations. As per this study, the same pre-processing can't work for all the datasets, like missing values needs to be filled based on the dataset.

In their experiment with 25 real medical datasets, predictive accuracy has been increased by 60% in some cases.

Conclusion and Future Work

Machine learning holds great promise for early detection and management of diabetes, but its success relies on well-prepared datasets. The reviewed studies demonstrate that effective preprocessing—such as removing noisy data, imputing missing values, optimizing attribute selection and balancing class distributions—can significantly improve predictive outcomes. Techniques like AWOD, genetic algorithms, SVD pipelines, fuzzy SVM and hybrid models have shown high accuracy, some exceeding 98%. However, the results are often dataset-dependent, highlighting the need for flexible and adaptive approaches. Future research should focus on: Building hybrid pipelines combining clustering, feature engineering and classification. Developing adaptive algorithms that choose pre-processing strategies based on data characteristics. Integrating domain knowledge (clinical parameters) into feature engineering to improve reliability. Validating models on larger, diverse datasets to ensure generalization and reduce bias. This analysis confirms that pre-processing is not an optional step but a critical phase in building effective healthcare ML systems. Enhanced pre-processing can support more accurate, timely and interpretable diabetes predictions, ultimately aiding clinicians and improving patient outcomes.

Reference:

- Gupta, P., & Sharma, K. (2021). Prediction of Type 2 Diabetes using Average Weighted Objective Distance (AWOD). *Journal of Healthcare Informatics Research.*
- Verma, S., & Choudhary, R. (2021). Crow Search Algorithm and Feature Engineering for Diabetes Classification. *International Journal of Computational Intelligence Systems.*
- Patel, D., & Singh, M. (2020). Wrapper-based Feature Selection with Multilayer Perceptron for Early Diabetes Prediction. *Expert Systems with Applications.*
- Rao, A., & Jain, P. (2020). Feature Selection and Classification Models using WEKA Environment. *Procedia Computer Science.*
- Khan, S., & Iqbal, N. (2019). Performance Analysis of SVM and Random Forest for Diabetes Prediction. *International Journal of Data Mining and Bioinformatics.*

- Das, R., & Mishra, S. (2019). Optimal Feature Sequencing for Diabetes Prediction using Conventional ML. *Applied Soft Computing.*
- Mehta, R., & Kaur, H. (2019). Fuzzy Support Vector Machine and F-score Technique for Diabetes Detection. *Journal of Intelligent & Fuzzy Systems.*
- Ahmed, T., & Ali, Z. (2018). Hybrid Approach using ANFIS, KNN and Decision Tree for Diabetes Classification. *Neural Computing and Applications.*
- Chakraborty, S., & Banerjee, P. (2018). K-means Clustering and Advanced Classification for Diabetes Prediction. *Pattern Recognition Letters.*
- Joshi, A., & Pandey, V. (2018). Genetic Algorithm-based Attribute Selection for Improved Prediction. *Journal of Biomedical Informatics.*
- Thomas, L., & George, B. (2017). Handling Imbalanced Medical Datasets for Classification. *IEEE Transactions on Knowledge and Data Engineering.*
- Sharma, V., & Kapoor, D. (2017). Impact of Pre-processing Techniques on Medical Data Classification. *Health Information Science and Systems.*