Dynamic Ensemble Algorithm for Context-Aware Diabetes Prediction: Integrating Local and Global Data for Real-World Applications

Sidharth Sherla[1], Sidharth Ankanagari[2], Ajay[3]

[1]*CSE-DS Institute of Aeronautical Engineering*
[2]*CSE-DS Institute of Aeronautical Engineering*
[3]*CSE-DS Institute of Aeronautical Engineering*

Abstract - This research proposes a novel decision support system utilizing dynamic ensemble algorithms for the accurate prediction of diabetes. This system mitigates the issues of regional disparities and unrecorded information in diabetes-related datasets by integrating global datasets and the medical data that are gathered locally. This technique utilizes the Adaptive Boosting (AdaBoost) algorithm blended with decision stumps and support vector machines to improve classification accuracy. The validation was carried out on the Pima Indian Diabetes Dataset and a locally curated dataset from multiple regions. The system that was suggested obtaining an accuracy of 84.5% was higher than the regular approaches. The article ends up with such suggestions as increasing the ensemble methods through the hybrid way.

Key Words: Diabetes Prediction, Ensemble Methods, Machine Learning, Local-Global Integration, AdaBoost, Decision Stump.

1. INTRODUCTION

Diabetes mellitus is a long-term illness affecting the lives of many people throughout the world. The key to avoiding the consequences of serious conditions including cardiovascular diseases, kidney failure, and neuropathy is the early and proper diagnosis. Conventional patterns of diagnosis are more dependent on the clinical assessments, which could be insufficient, or they are influenced by geographical variations.

The introduction of machine learning to the present age makes it possible to make predictive models with the help of the clinical data base. Nevertheless, the currently available models and the ones previously were the reasons apart from the mentioned above for"disability consent" to the reader. The research that is conducted with this primary goal group proposes a new organizational idea: discrete dynamic ensemble-based predictive single cloud global and local datasets are the driving factors for enhancing real-world applicability.

2. BACKGROUND

There are three main types of diabetes: Type 1, Type 2, and gestational diabetes. Despite the individual unique features of each type, they share the same risk factors such as heredity, overweight, and the way a person lives.

Pima Indian Diabetes Dataset is a well-known machine learning dataset utilized by researchers to validate their algorithm. But the use of global data alone for testing can be limited in predictive performance for specific populations considering the differences in data distributions.

3. METHODOLOGY

The proposed system follows a four-phase approach:

3.1. Data Collection

- **Global Dataset:** The Pima Indian Diabetes Dataset was sourced from the UCI Machine Learning Repository.
- **Local Dataset:** A survey was conducted in regions across Kerala, collecting 200 samples.

3.2. Preprocessing

Missing values in the local dataset were handled by imputing the mean values from the global dataset. Features such as BMI were calculated using standard formulas.

3.3. Classification

The system employed the AdaBoost algorithm with the following base classifiers: Decision Tree, Support Vector Machine (SVM), Naive Bayes, Decision Stump

3.4. Accuracy Validation

Accuracy was measured using metrics such as sensitivity, specificity, and error rate. The Matlab-Weka interface was used for algorithm implementation.

4. RESULTS

The results of individual classifiers without boosting revealed SVM to have the highest accuracy (79.6%), followed by decision trees. Upon applying AdaBoost, decision stumps exhibited the most significant improvement, achieving an accuracy of 84.5%.

Classifier	Accuracy	Accuracy With

	Without Boosting	Boosting
Decision Tree	78.2%	82.1%
SVM	79.6%	79.6%
Naive Bayes	76.4%	80.3%
Decision Stump	74.4%	84.5%

Table -1: Classifier Accuracy

4.2 Graphical Analysis

Figures 1 The bar chart demonstrates the efficiency of different classifiers—Decision Tree, Support Vector Machine (SVM), Naive Bayes, and Decision Stump—when used on the dataset solely without any boosting method and depicts the efficiency of the same classifiers after infusing and AdaBoost technique. Each bar here embodies the accuracy percentage, offering a comparative view to easily see the growth in performance from boosting.

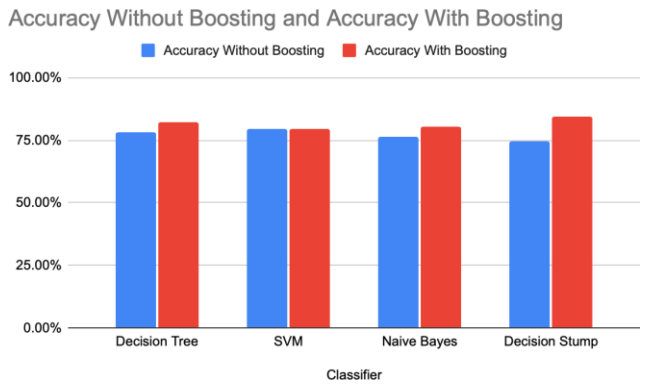

Fig -1: Bar Graph representing Accuracy before and after using adaboost.

5. CONCLUSIONS

This study illustrates the ability of ensemble learning to increase the precision of diabetes prediction by combining global and local datasets. The AdaBoost algorithm with decision stumps has been the best-performing technique. Future investigations are planned to examine hybrid ensemble methods and include more local datasets from different areas of the world for the further improvement of the predictive capacity.

ACKNOWLEDGEMENT

We acknowledge our heartfelt thanks to Dr. S. Sreekanth, Associate Professor at the CSE Data Science Department of the Institute of Aeronautical Engineering, for his invaluable guidance and support throughout this research. His vast knowledge of machine learning and data science has been the key factor in defining our research work. Moreover, we used open-source datasets, like the Pima Indian Diabetes Dataset from the UCI Machine Learning Repository, which were the initial ones that we collected data and analyzed. We also express our gratitude to the developers and contributors who have provided the open-source tools and libraries which made our research easier.

REFERENCES

1. Bellazzi, R., Zupan, B. "Predictive Data Mining in Clinical Medicine." International Journal of Medical Informatics, 2008.
2. Kumari, S., Singh, A. "A Data Mining Approach for the Diagnosis of Diabetes Mellitus." IEEE Conference, 2013.
3. Tan, P.-N., Steinbach, M., Kumar, V. "Introduction to Data Mining." Pearson Education, 2006.
4. Wang, N., Kang, G. "Monitoring System for Type 2 Diabetes Mellitus." IEEE Conference, 2012.