

Chapter 1


A Comprehensive Approach for Using Hybrid Ensemble Methods for Diabetes Detection

Md Sakir Ahmed

 <https://orcid.org/0000-0002-7754-639X>

Department of Computer Applications, Assam Don Bosco University, India

Abhijit Bora

 <https://orcid.org/0009-0009-7481-0835>

Assam Don Bosco University, India

ABSTRACT

This study is focused on the possible application of hybrid models as well as their usage in the detection of diabetes. This study focuses on various machine learning algorithms like Decision Trees, Random Forests, Logistic Regression, K-nearest neighbor, Support Vector Machines, Gaussian Naive Bayes, Adaptive Boosting Classifier, and Extreme Gradient Boosting as well as the usage of Stacking Classifier for the preparation of the hybrid model. An in-depth analysis was also made during this study to compare the traditional approach with the hybrid approach. Moreover, the usage of data augmentation and its application during an analysis has also been discussed along with the application of hyperparameter tuning and cross-validation during training of the various models.

DOI: 10.4018/979-8-3693-2260-4.ch001

1. INTRODUCTION

1.1 Background of the problem

With the increase in the consumption of processed foods, there has also been an increase in the number of cases of diabetes. There has been a linear increase in the number of cases, affecting a vast spectrum of the global population ranging from children to adults to seasoned citizens. This sudden increase can be directly correlated to increased consumption of processed foods as per studies, however, this is not the only factor. Lack of physical activity, consumption of alcohol, smoking, and improper sleep schedules are some of the contributing factors to the rise in the number of cases. Traditional approaches for diabetes detection include urine tests, random blood sugar tests, clinical symptoms, risk assessments, etc. However, with the advent of technology emphasis needs to be given to finding newer methods to identify and assess the various risk factors as well as for early detection of diabetes. This may in turn reduce the number of cases occurring annually enabling a healthier life for the global population.

1.2 Proposed solution

This study is a brief introduction to hybrid ensemble learning models and focuses on giving a detailed overview of the possible implications of these models on early diagnosis as well as their usage for the identification of risk factors. Several machine learning algorithms like Decision Trees, Random Forest, Logistic Regression, K-nearest neighbors, Support Vector Machines, Gaussian Naïve Bayes, and Adaptive Boosting Classifier can be used to diagnose diabetes as well as other diseases quite accurately. The accuracy can be further increased by stacking multiple models using a stacking classifier. These stacked models are known as hybrid models and provide better accuracy compared to just using a single classifier due to their robustness in identifying noise in the data, better hyperparameter tuning, enhanced adaptability, and improved generalization.

In this study, several traditional machine-learning algorithms were used along with hyperparameter tuning to find the best suitable parameter for the given data after which the best-performing models were stacked together along with Extreme Gradient Boosting Classifier, and their accuracy for training and testing was calculated. The observations are discussed in detail in Sections 3 and 4.

13 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage: www.igi-global.com/chapter/a-comprehensive-approach-for-using-hybrid-ensemble-methods-for-diabetes-detection/343879

Related Content

Bitmap Indices for Data Warehouses

Kurt Stockinger and Kesheng Wu (2007). *Data Warehouses and OLAP: Concepts, Architectures and Solutions* (pp. 157-178).

www.irma-international.org/chapter/bitmap-indices-data-warehouses/7620

Marketing Data Mining

Victor S.Y. Lo (2008). *Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications* (pp. 2824-2832).

www.irma-international.org/chapter/marketing-data-mining/7803

Strategic Utilization of Data Mining

Chandra S. Amaravadi (2008). *Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications* (pp. 1689-1695).

www.irma-international.org/chapter/strategic-utilization-data-mining/7724

A Single Pass Algorithm for Discovering Significant Intervals in Time-Series Data

Sagar Savla and Sharma Chakravarthy (2008). *Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications* (pp. 3272-3284).

www.irma-international.org/chapter/single-pass-algorithm-discovering-significant/7833

Classification Methods

Aijun An (2005). *Encyclopedia of Data Warehousing and Mining* (pp. 144-149).

www.irma-international.org/chapter/classification-methods/10582