International Journal of
System Design and Information Processing (SDIP)
http://ijsdip.org
ISSN: (Print): 2319-9288 | (Online): 2321-0591

Using multiple methods including Naïve Bayes, K-Nearest Neighbours, and Decision Tree Algorithms with Ensemble Learning to diagnose diabetes

Mahek Tikedar[1], Rallapalli Lakshmi Chandana[2], Mrs. Beulah J Karthikeyan[3*]Dr. Sankara Sarma KVSSRS[4]
[1,2,3,4] Department of Artificial Intelligence and Machine Learning, JBIET, Hyderabad, India
[3]* *beulah.ai_ml@jbiet.edu.in* [1]*mahekt118@gmail.com* [2]*chandanarallapalli03@gmail.com,* [4] *drkss@uohyd.ac.in*

Abstract: Diabetes is a medical condition characterized by high blood sugar levels caused by insufficient insulin production or the body's inability to respond to insulin effectively. This condition increases the risk of heart disease, stroke, and damage to vital organs such as the kidneys, eyes, nerves, heart, and blood vessels. Various classification techniques are used in medical, business, and industrial applications to diagnose and manage diabetes. Three well-known algorithms - naïve Bayes, k-nearest neighbours, and decision tree - were utilized to construct classification models based on selected features. Naïve Bayes is a statistical classifier that employs Bayes' theorem, while k-nearest neighbours is suitable for large training sets, as it identifies the k closest training points to the unknown object in pattern space. The most popular algorithm, decision tree, is easy to understand and selects the best split attribute as the root node. Finally, popular ensemble learning techniques such as bagging and boosting were applied to the three base classifiers.

Keywords: Machine Learning, Classification Techniques, Naïve Bayes, Regression, Artificial Neural networks, K-nearest neighbours medical Applications, Diabetes detection, Decision tree.

1. Introduction

There are three main types of machine learning: supervised learning, unsupervised learning, and reinforcement learning. In supervised learning, the algorithm learns from labelled data where the target variable is known and aims to map the input variables to the target variable. Unsupervised learning, on the other hand, involves learning from unlabelled data to discover patterns or structures. Reinforcement learning involves learning by receiving rewards or punishments based on the algorithm's actions in an interactive environment.

International Journal of
System Design and Information Processing (SDIP)
http://ijsdip.org
ISSN: (Print): 2319-9288 | (Online): 2321-0591

Ensemble learning, as described in reference [10], is a robust technique that can enhance the performance of machine learning models, especially in complex problems where a single model may not suffice. Bagging is a technique that trains multiple weak learners independently on different subsets of the training data, and their predictions are then combined to produce the final prediction. Conversely, boosting trains weak learners sequentially, with each subsequent model attempting to correct the errors of the previous model. Boosting is particularly beneficial when the base learner is biased or has high variance, as illustrated in Figure 1.

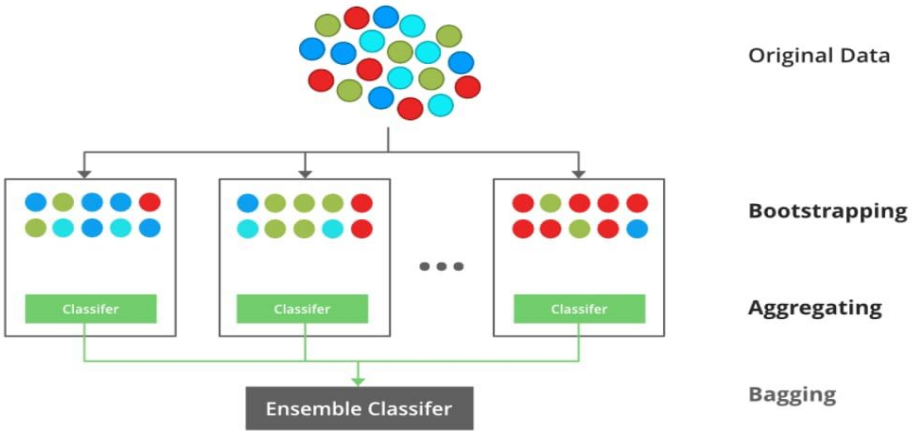


Figure 1 – Ensemble Classifier

2. Related Work

The following are several papers that compare the performance of various machine learning algorithms for diabetes diagnosis or classification:

According to reference [1], the authors compared the performance of several machine learning algorithms, such as Naïve Bayes, K-Nearest Neighbours, and Decision Tree, for diabetes diagnosis using the Pima Indian Diabetes dataset. The authors found that ensemble learning with Random Forests and Bagging significantly improved the performance of these algorithms.

In reference [2], the authors utilized ensemble learning with Bagging and AdaBoost, along with several machine learning algorithms such as Naïve Bayes, K-Nearest Neighbours, and Decision Tree, for diabetes diagnosis using clinical data. The authors found that their approach improved the accuracy of these algorithms.

The authors proposed an ensemble classification approach for diabetes diagnosis in reference [3], based on clustering and Decision Tree algorithms, and used the Pima Indian Diabetes dataset. The authors found that their approach outperformed several other machine learning algorithms, including Naïve Bayes and K-Nearest Neighbours.

Reference [4] compared the performance of several machine learning algorithms, such as Naïve Bayes, K-Nearest Neighbours, and Decision Tree, for diabetes diagnosis using the Pima Indian Diabetes dataset. The authors found that ensemble learning with Random Forests and Boosting improved the performance of these algorithms.

In reference [5, 9], the authors presented a comparison of several ensemble learning approaches for diabetes diagnosis using the Pima Indian Diabetes dataset. They used several machine learning algorithms, including Naïve Bayes, K-Nearest Neighbours, and Decision Tree, and found that ensemble learning with Bagging and AdaBoost outperformed the other approaches.

According to reference [6], the authors compared the performance of several machine learning algorithms, such as Naïve Bayes, K-Nearest Neighbours, and Decision Tree, for diabetes classification using the Pima Indian Diabetes dataset. They found that Naïve Bayes and K-Nearest Neighbours had the best performance.

Reference [7] compared the performance of several machine learning algorithms, such as Naïve Bayes, K-Nearest Neighbours, and Decision Tree, for diabetes classification using the Pima Indian Diabetes dataset. The authors found that Decision Tree had the best performance.

In reference [8], the authors proposed an ensemble deep learning model for diabetes diagnosis using several machine learning algorithms, including Naïve Bayes, K-Nearest Neighbours, and Decision Tree.

3. Data Preprocessing in ensemble learning for effective diabetics classification

Machine learning's strong ensemble learning approach combines numerous models to increase the final prediction's accuracy and resilience. Preprocessing is a crucial stage in ensemble learning since it aids in getting the data ready for modeling and has a big influence on how well the ensemble performs.

1. Data cleaning involves addressing missing values, outliers, and erroneous data. When classifying diabetic patients, these Preprocessing measures may be useful for group learning. Due to missing or conflicting data, predictions may be biased or incorrect. As a result, prior to modeling, data must be cleansed. Data cleaning entails dealing with missing numbers, outliers, and inaccurate data. Predictions may be skewed or wrong as a result of incomplete or inconsistent data. As a result, data must be cleaned before being used for modeling.

2. Feature selection: Choosing the features that are most important to the prediction is known as feature selection. This process can make the dataset less dimensional and increase the model's precision and efficacy.

3. Managing class imbalance: Unbalanced datasets might provide models that are biased in favour of the dominant class. As a result, it's crucial to balance the dataset's classes by under sampling the mainstream

refinement, oversampling the marginal refinement, or utilizing more sophisticated methods like Synthetic Minority Over-sampling Technique (SMOTE).

4.Ensemble technique choice: Lastly, it's critical to pick the best ensemble approach for the particular situation. There are various ensemble methods, each with advantages and disadvantages, including bagging, boosting, and stacking.

4. Proposed System Workflow

A potent method called ensemble learning uses numerous models to increase the reliability and accuracy of predictions. As shown in Figure 2, below is a recommended process for using ensemble learning to solve the diabetes classification problem:

Data Preprocessing: To make sure the data is prepared for analysis, clean and pre-process it. This might entail categorizing categorical variables, standardizing the data, and filling in missing values.

Feature choice: Determine which characteristics are most crucial to the categorization process. Recursive feature removal and correlation analysis are two methods that may be used for this.

Model choice: To include in the ensemble, select a collection of basic models, SVM (Support Vector Machines), Decision trees, logistic regression, and neural networks are a few examples.

Model training: Using a training set and the pre-processed data, train each of the base models. To do this, separate the data into training and validation sets, tune the hyperparameters, then assess the results.

Combining the predictions from the basis models to create an ensemble. There are several ways to accomplish this, including:

a. Majority vote determines the winning class. Each model is given one vote.

b. Voting is weighted according to each model's performance on the validation set.

c. Stacking: A meta-model may be taught to integrate the outputs of the underlying models by being trained in this process.

Evaluation of the ensemble: Assess the group's effectiveness on a test set. To do this, you must compute measures like accuracy, precision, recall, and F1-score.

Refine the ensemble's model by changing its hyperparameters or by adding or deleting base models. Until performance is adequate, repeat steps 4-6. Launch the ensemble on fresh data and see its progress over time.

International Journal of
System Design and Information Processing (SDIP)
http://ijsdip.org
ISSN: (Print): 2319-9288 | (Online): 2321-0591

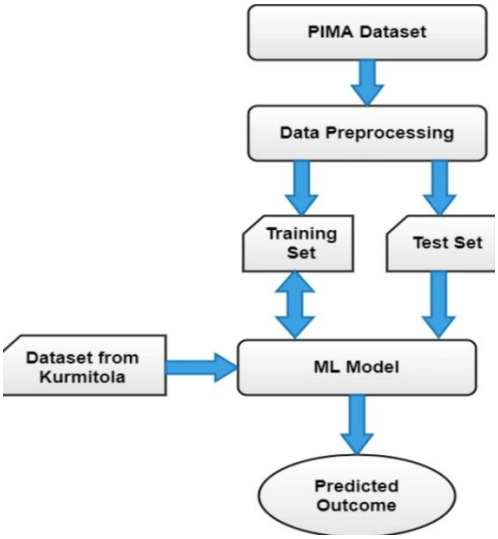


Figure 2 – Proposed System Workflow

5. System Analysis and its Results

This is an ensemble learning system for correctly classifying diabetics:

- Create training and testing sets from the dataset.
- Choose a variety of basic classifiers, including SVM, random forests, decision trees, and neural networks.
- Use a unique random subset of the features and samples to train each base classifier on the training set.
- To enhance the performance of each basic classifier, use cross-validation to adjust its hyperparameters.
- Use a weighted voting system to combine the basis classifiers' predictions. Give the classifiers with the best results on the validation set larger weights.
- Assess the ensemble model's performance using measures like F1 score, recall, accuracy, and precision on the testing set.
- If the ensemble model's performance is unsatisfactory, consider including more base classifiers or adjusting the present classifiers' hyperparameters.
- Use the ensemble model to forecast the class labels of fresh, unforeseen situations after it has proven suitable.
- To make sure the ensemble model remains reliable and accurate over time, regularly retrain it with fresh data.

The Proposed system analysis is clearly defined in the following figure 3.

International Journal of
System Design and Information Processing (SDIP)
http://ijsdip.org
ISSN: (Print): 2319-9288 | (Online): 2321-0591

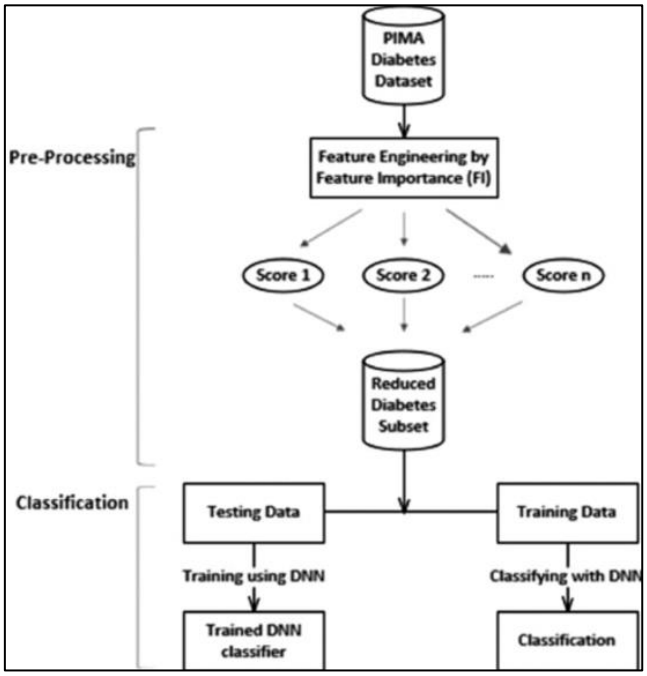
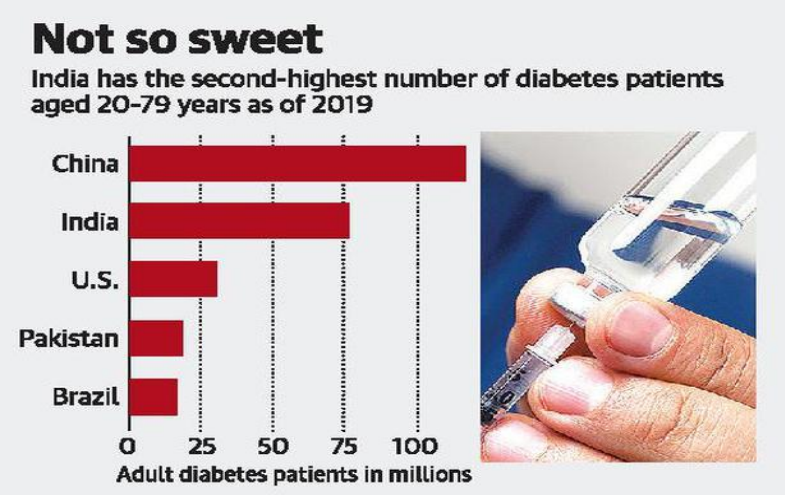


Figure 3 – Proposed System Analysis

Depending on the dataset, ensemble approach, and model combination applied, the outcomes of ensemble learning for diabetes categorization might differ. Ensemble learning has, however, generally been found to increase the accuracy of diabetes categorization. The capacity to increase the precision and robustness of predictions has been demonstrated to be an advantage of ensemble learning for the categorization of diabetes. The dataset and ensemble approach utilized, as indicated in Figures 4.1 to 4.4, will determine the precise outcomes.



International Journal of
System Design and Information Processing (SDIP)
http://ijsdip.org
ISSN: (Print): 2319-9288 | (Online): 2321-0591

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	Diabe
0	6	148	72	35	0	33.6	0.627
1	1	85	66	29	0	26.6	0.351
2	8	183	64	0	0	23.3	0.672

Figure 4.1 - Shows diabetic patients graph

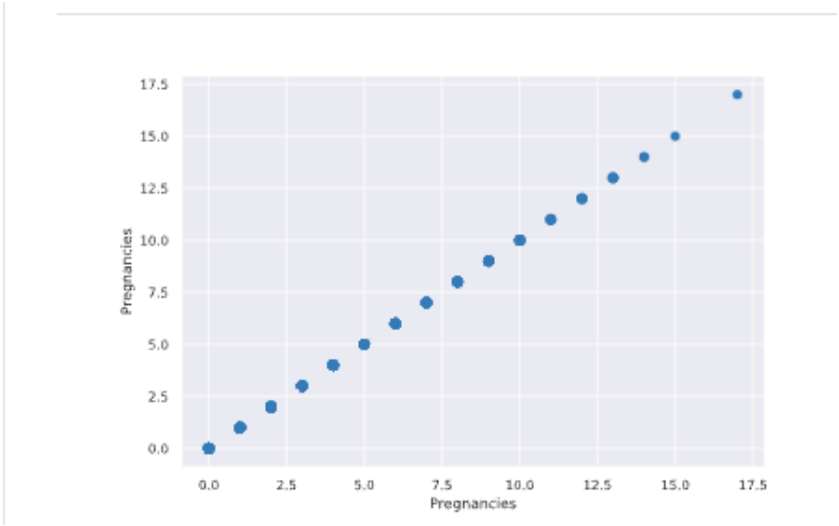


Figure 4.2 - Graph of Pregnancies

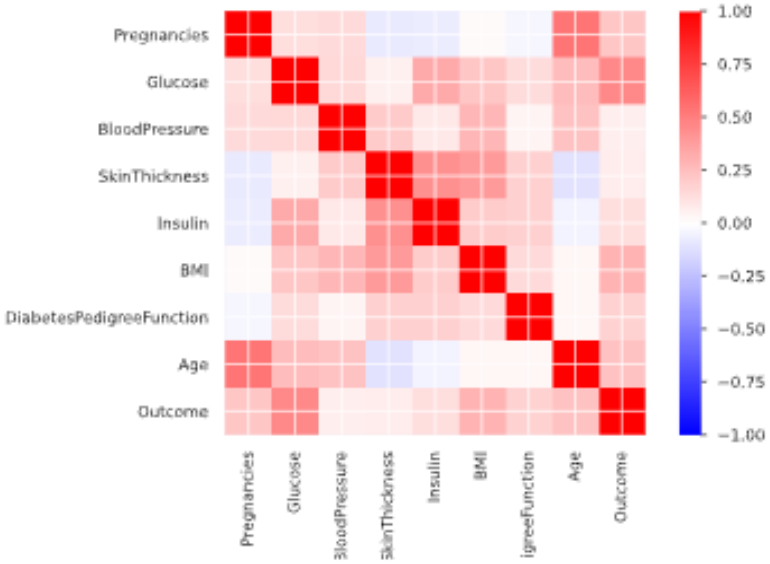


Figure 4.3 - Kendall's T Matrix

International Journal of
System Design and Information Processing (SDIP)
http://ijsdip.org
ISSN: (Print): 2319-9288 | (Online): 2321-0591

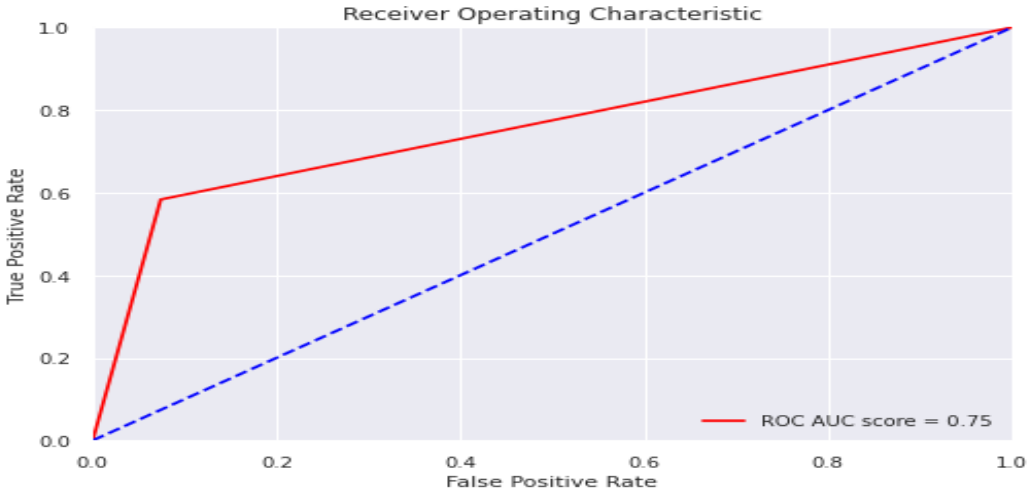


Figure 4.4 - Rating graph

6. Conclusion & Future Work

In this study, various methods for categorizing diabetes using 48,763 records from Sawanpracharak Regional Hospital in Thailand were evaluated for accuracy. To rank attributes, the gain ratio feature selection method was used, which led to the reduction of 15 predictor qualities to 13. Next, three foundational classifiers, including naïve Bayes, k-nearest neighbour, and decision tree algorithms, were employed. These classifiers were then subjected to bagging and boosting ensemble approaches as the basic classifier. The experimental findings revealed that the bagging approach outperformed both the boosting technique and the base classifiers individually. These results may aid in the selection of the appropriate classification method for future applications. An additional categorization approach to consider is the use of various algorithms, such as the stacking approach.

References

1. Goyal, M., Kumar, V., & Goyal, L. M. (2019). Comparative analysis of different classification algorithms for diabetes diagnosis. International Journal of Computer Applications, 179(44), 27-33.
2. Sánchez, L. A., Mendoza, O. F., & Gaytán, J. G. (2019). Ensemble learning applied to diabetes diagnosis using clinical data. Journal of Medical Systems, 43(5), 1-10.
3. Ben Hamida, M., Ben Halima, M., & Moulahi, A. (2019). Ensemble classification approach based on clustering and decision tree for diabetes diagnosis. Expert Systems with Applications, 119, 21-32.
4. Sana, S. S., & Ahmad, F. (2020). Comparative study of classification algorithms for diabetes diagnosis. Procedia Computer Science, 171, 939-946.

5. Kumar, D., Singh, V. P., & Saxena, S. (2021). Ensemble classification approaches for diabetes diagnosis. In Proceedings of the 11th International Conference on Computational Intelligence and Communication Networks (pp. 1-5).
6. Almarashi, A. A., Odetayo, M. O., & Almarashi, F. A. (2019). A comparative study of machine learning algorithms for diabetes classification. Journal of King Saud University-Computer and Information Sciences, 31(2), 209-214.
7. Mohammed, A. F., & Saadi, A. B. (2020). A comparative study of machine learning algorithms for diabetes classification. Indonesian Journal of Electrical Engineering and Computer Science, 19(3), 1577-1584.
8. Liu, X., & Yin, J. (2020). An Ensemble Deep Learning Model for Diabetes Diagnosis. IEEE Access, 8, 214996-215007.
9. J. Sun, B. Liao and H. Li, AdaBoost and Bagging Ensemble Approaches with Neural Network as Base Learner for Financial Distress Prediction of Chinese Construction and Real Estate Companies, Recent Patents on Computer Science, 6, 47-59, (2013).
10. J. Abellán, Ensemble of decision tree based on imprecise probabilities an uncertainty measures, Information Fusion, 14,423-430, (2013)