

Implementation of a Web Application to Foresee and Pretreat Diabetes Mellitus in Women using Machine Learning

Nishitha Akula¹, Sonu Sagar², Swathi Sridharan³

¹VIII Semester, Dept. of ISE, BNMIT

²VIII Semester, Dept. of ISE, BNMIT

³Assistant Professor, Dept. of ISE, BNMIT, Karnataka, India

Abstract -According to the International Diabetes Federation (IDF), there are 537 million people living with diabetes as of 2021. The numbers are projected to rise to 643 million by 2030 and 783 million by 2045. Almost 1 in 2 (240 million) adults living with diabetes are undiagnosed. 1 in 6 live births (21 million) are affected by diabetes during pregnancy.

Since people with diabetes are at risk from further complications such as neuropathy, retinopathy, stroke, kidney diseases, amputation etc. implementing a system that can predict diabetes early on is necessary.

The most recent advances in machine learning and data mining technologies can be used to uncover hidden patterns, which may aid in the early detection and effective pre-treatment of diabetes. Seven popular machine learning algorithms like Logistic Regression, Decision Tree, Random Forest, SVM, KNN, Adaboost and Gradient Boost were tested. The top five high performing machine learning algorithms were considered and used in Ensemble Voting Classifier.

The suggested framework works better than the individual algorithms discussed in the paper with an accuracy of around 82%. Further, the proposed system also sends recommendations in the form of diet charts and exercises for pre-treatment.

Key Words: Diabetes prediction, Ensemble methods, Machine Learning, Pima Indians Diabetes dataset, Diabetes pre-treatment, Web application

1. INTRODUCTION

Diabetes mellitus, also known as diabetes, is an incurable chronic disease caused due to deficiency of a hormone called insulin [1]. Insulin produced by the pancreas in the body, allows glucose from food to enter the bloodstream. Diabetes occurs when the pancreas malfunctions, resulting in coma, pathological destruction of pancreatic beta cells, sexual dysfunction, renal and retinal failure, weight loss, cerebral vascular dysfunction, ulcer, cardiovascular dysfunction, joint failure, pathogenic effects on immunity and peripheral vascular diseases [2].

Diabetes is classified into two types: type 1 and type 2. Diabetes type 1 accounts for 5 to 10% of all cases of diabetes.

This form of diabetes is most prevalent in children or adolescents and is distinguished by a limited function of the pancreas. Since the pancreas remains partially functional, type 1 diabetes does not cause symptoms initially. The disease is not visible until 80-90% of pancreatic insulin-producing cells have been destroyed [3].

Diabetes type 2 accounts for 90% of the overall cases of diabetes. Chronic hyperglycemia and the body's inability to maintain blood sugar levels distinguish this type of diabetes, resulting in an unusually elevated level of glucose in the blood. [4].

Recent healthcare studies have used a wide range of technologies to diagnose patients and predict their disease based on clinical data. ML techniques are now being used in the healthcare system to more accurately predict diabetes. ML allows a computer to learn from experiences or inputs (such as clinical data) and predict the output category (existence of disease). ML techniques are classified as supervised, unsupervised, or reinforcement learning. The features, as well as the target class, are used as input for learning in supervised learning. The input data is provided without the target class in unsupervised learning. This datapoints are then clustered using a similarity metric. Reinforcement learning is a method that employs the hit-and-trial method to determine the best outcome. Award and penalty values are used to achieve the best results.

The primary goal of this research is to propose the development of an improved prognostic tool for early diabetes prediction and pre-treatment. A large amount of data and datasets are available on the web or from independent bodies. The PIMA Indians diabetes dataset used in this work is one of the most commonly used datasets in several studies, and it is gathered by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). Different ML classifiers (Logistic Regression, Decision Tree, Random Forest, SVM, KNN, Adaboost and Gradient Boost) were implemented in our proposed framework. Furthermore, we propose an ensembling classifier using a number of base models for improving diabetes prediction. Hard weighted voting is used to ensemble the ML models.

The remaining part of the paper is organized as follows. In Section 2, we discuss some relevant work that was done

previously. Section 3 elaborates on the proposed architecture and various algorithms tested in this study. Section 4 examines the results obtained and the final implementation in an end-to-end webapp. In Section 5, we conclude the paper by presenting future enhancements.

2. LITERATURE REVIEW

In recent years, numerous models have been proposed and published. The authors of [5] presented a theoretical approach based on three classification techniques: Support Vector Machines, Logistic Regression, and Artificial Neural Network.

On the Pima Indians Diabetes Dataset, Sisodia and Sisodia [6] used classification algorithms such as decision trees, Naive Bayes and support vector machine, with the Naive Bayes classifier achieving the highest accuracy in diabetes prediction. Sisodia used a tenfold cross-validation technique. In this technique, the dataset was divided into ten equal parts, nine parts were used for training and the rest for testing. Diabetes was predicted using evaluation parameters such as recall, accuracy, area under the curve and precision

Kumari et al. [7] implemented random forest, Naive Bayes and logistic regression to the Pima Indians Diabetes Dataset and compared these to the ensemble approach. The model outperformed the ensemble classifier with an accuracy of 79%.

Kandhasamy and Balamurali [8] applied machine learning algorithms such as random forest, J48, support vector machine and k-nearest neighbours to predict diabetes using the dataset in the UCI repository. The authors used the aforementioned classifier twice, once before and once after pre-processing the data. The techniques for pre-processing were not described, except that the dataset contained some noise that was removed later. The authors assessed the prediction based on its accuracy, sensitivity and specificity. The decision tree achieved the highest accuracy of 73.82% when the data was not pre-processed. With pre-processing, random forest obtained an accuracy of 100%.

Perveen et al. [9] used data from the Canadian primary care sentinel surveillance network in their study. The dataset contained the following attributes: gender, BMI, fasting blood sugar, triglycerides, systolic blood pressure, and diastolic blood pressure. The authors implemented the bootstrap, decision trees, and adaptive boosting classifiers.

The performance of various classification techniques [10,11] such as Support Vector Machine, Decision Stump, Decision Trees and Naive Bayes without boosting was evaluated and gave an accuracy of 79.68%, 74.47%, 76% and 79.68% respectively. The performance evaluation of the above algorithms with Adaboost as the base classifier improved the accuracy except for support vector machines.

3. MATERIALS AND METHODS

This section sheds light on the materials and techniques implemented in the experiment discussed in the paper.

3.1 Dataset

The ML models were trained and tested using the PIMA Indians Diabetes dataset, which contained the details of 768 female diabetic patients from the Pima Indian community near Phoenix, Arizona [12]. This dataset comprises of 268 diabetic (positive) and 500 non-diabetic (negative) patients with eight distinct characteristics. The attributes are as follows: Number of pregnancies, Glucose in plasma, Insulin level, Diabetes pedigree function, Skin thickness, Body mass index and Age. The diabetes pedigree function was computed [12] as in (1).

$$\text{Pedigree} = \frac{\sum_i K_i (88 - ADM_i) + 20}{\sum_j K_j (ALC_j - 14) + 50} \quad (1)$$

where i and j represent the relatives who had developed and not developed diabetes respectively. K refers to the proportion of genes shared with relatives ($K = 0.500$ for a parent or full sibling, $K = 0.125$ for a half aunt, half-uncle, or first cousin and $K = 0.250$ for a half-sibling, grandparent, aunt or uncle). ADM_i and ACL_j denote the ages of relatives at the time of diagnosis and the most recent non-diabetic test.

3.2 Proposed framework

The construction of algorithms capable of producing general patterns and hypotheses by using supplied instances to predict the outcome of future instances is known as supervised machine learning. The goal of supervised machine learning classification algorithms is to categorise data based on prior knowledge. They learn the pattern from previous data and attempt to predict new results. ML algorithms such as rule-based, instance-based, function-based, probability-based, tree-based, and so on are used to locate available data. The architecture of our proposed model is shown in Fig.1.

3.3 Preprocessing

Pre-processing included the following steps:

1. Removing outliers
2. Checking for target imbalance
3. Checking for missing values. No missing values were found in the dataset.
4. In case the value of few fields was zero, they were replaced with the mean value of the respective attribute. This is necessary to obtain normal distribution.

- Despite this, there were some attributes that did not have normal distribution. Non-linear scaling was thus performed on them.

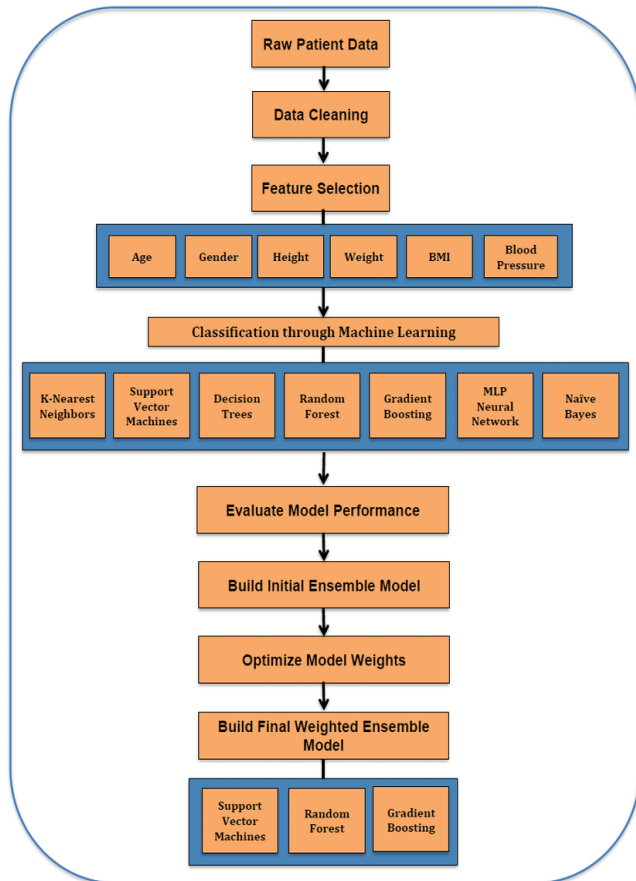


Figure -1: The proposed architecture diagram for diabetes prediction

3.4 Algorithms tested

The following seven algorithms were executed and evaluated in this study.

Logistic Regression

Based on a set of independent variables, logistic regression calculates the likelihood of an event occurring, such as True or False. The dependent variable has a range of 0 to 1 because the outcome is a probability. A logit transformation is applied to the odds in logistic regression, which is the probability of success divided by the probability of failure. This is also referred to as log odds or the natural logarithm of odds. The log likelihood function is produced by all of these iterations, and logistic regression attempts to maximise this function to find the best parameter estimate. Once the optimal coefficient (or coefficients, if more than one independent variable is present) has been determined, the conditional probabilities for each observation can be calculated, logged, and summed to produce a predicted

probability. A probability less than .5 predicts 0 and a probability greater than 0 predicts 1.

Naïve Bayes

The Bayes Theorem is the basis of the Naive Bayes algorithm, which is used in a broad array of classification tasks. The Bayes' Theorem is a mathematical equation used to determine conditional probabilities. Conditional probability is an estimate of the likelihood of one event occurring given that another event has already occurred (either through assumption, presumption, assertion, or evidence). Given the class, Naïve Bayes works with the assumption that all the attributes are conditionally independent.

Decision Trees

A decision tree is a graph in which the internal nodes serve as tests on the input data and the leaf nodes serve as classes of the input data. These tests are filtered down through the tree to find the appropriate output pattern for the input. The decision tree is applied in decision analysis to illustrate the output as a splitting rule for each individual attribute. It is a branching graph that is used to predict decision-making outcomes both visually and explicitly. Each attribute is regarded as a branching node, and at the end of the branch, a rule is constructed that divides values belonging to diverse classes. As the name implies, it is a tree-like structure that ends with a decision known as the leaf. The root is the most potentially useful attribute for predicting rule formation outcomes. Decision trees are simple and easy to implement, and it also predicts the results more accurately.

Random Forest

Random forest is a supervised classification algorithm that is widely used in classification tasks. It is composed of a series of decision trees. These trees have the same number of nodes but differ in their data. The outcomes of these various decision trees will be aggregated to produce a final result that indicates the average response of all of the decision trees. As the number of trees in a forest increase, the generalisation error approaches a limit. The generalisation error of a forest of tree classifiers is determined by the strength of the individual trees in the forest as well as their correlation.

Support Vector Machine

Support vector machines have shown to be extremely effective in a variety of classification tasks. It seeks the best separating hyperplane between classes by locating the coordinates on the edges of the class descriptors. The distance between the classes is represented by the margin. SVM algorithms find a margin with the greatest possible distance. SVMs are intended to handle binary classification data that can be separated linearly.

AdaBoost

AdaBoost (Adaptive Boosting), is a Machine Learning technique that is used as an Ensemble Method. The most common AdaBoost algorithm is decision trees with one level, which means decision trees with only one split. Because these trees are so short and only have one classification decision, they are often referred to as decision stumps. It is employed to enhance the effectiveness of any machine learning algorithm. It works best with slow learners. On a classification problem, these are the models that achieve accuracy just above random chance. Here, a series of weak classifiers are connected, with every weak classifier attempting to strengthen the classification of samples misclassified by the previous weak classifier. Boosting does this by combining weak classifiers in succession to construct a strong classifier.

Gradient Boosting

Gradient boosting is a type of boosting method that iteratively learns from each weak learner to build a strong model. It can improve: regression, classification and ranking. The word Gradient refers to the fact that multiple derivations of the same function is obtained. Gradient Boosting is a functional gradient iterative algorithm that reduces a loss function by recursively choosing a function that points in the direction of the negative gradient, also called a weak hypothesis.

4. RESULTS

The Pima Indians dataset considered was split in a 70 to 30 ratio for training and testing respectively. A thorough evaluation and performance analysis was performed on the seven algorithms (Logistic Regression, Decision Tree, Random Forest, SVM, KNN, AdaBoost and Gradient Boost) tested. The results are shown in Table 1.

Algorithms	Accuracy in %
Logistic Regression	79.7
Random Forest	77.1
Adaptive Boost	79.2
Gradient Boost	80.2
K Nearest Neighbour	76.6
Decision Tree	74.5
Support Vector	67.7

Table -1: Performance analysis of the algorithms tested

The top five highest performing algorithms i.e Logistic Regression, Random Forest, K-Nearest Neighbours, Adaboost and Gradient Boosting were combined to form an ensemble

model. Ensembling is a type of machine learning technique that aggregates numerous base models to produce a single robust model.

The hard voting classifier was implemented. Hard voting is an ensembling method that selects the class with the highest number of votes. The votes refer to each individual base model or algorithm in the ensemble. Chart 1 compares the performance of the five algorithms considered against the final ensemble model.

Ensemble models have a number of advantages. They outperform individual models in terms of predictive accuracy. They are very useful when the dataset contains both linear and non-linear data since different models can be combined to handle this type of data. With ensemble methods, bias or variance can be reduced, and the model is usually neither underfitted nor overfitted. Most importantly, ensemble models are always less noisy and more stable.

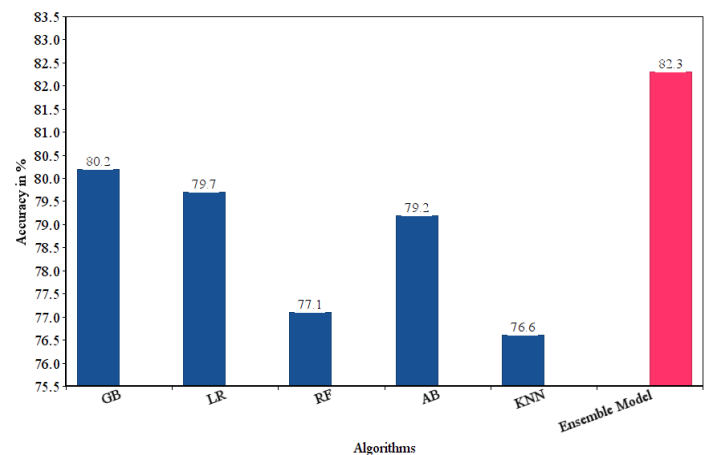


Chart -1: Ensemble model v/s Base Models

The final ensemble model built formed the crux of the end-to-end application that was created. A webapp using HTML, CSS and JavaScript was developed. The website would collect user details through a form in the input page as shown in Figure 2.

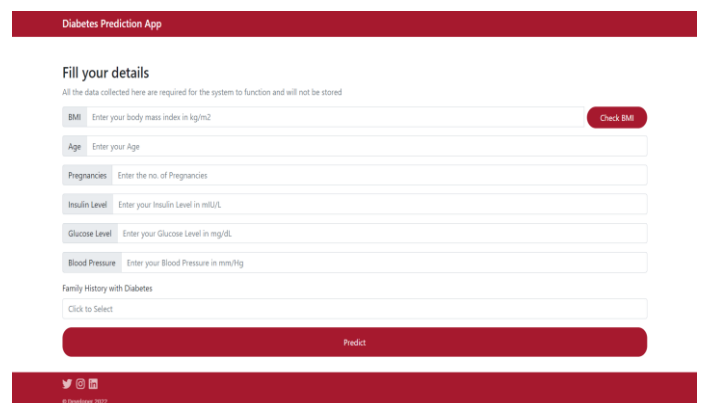


Figure -2: Input page of the final application

The model would run on the biological details provided and predict whether the individual is at a risk of getting diabetes. In case an individual is prone, a new page with recommended diet charts, exercises etc. are provided in order to pre-treat it effectively. This is shown in Figures 3 and 4.

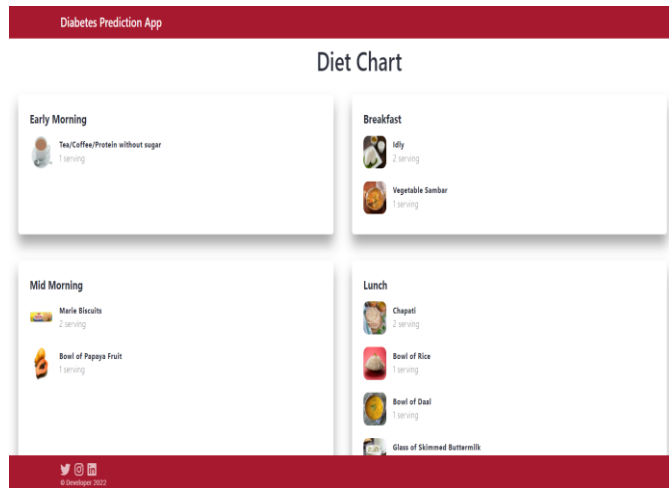


Figure -3: Diet charts for pre-treatment

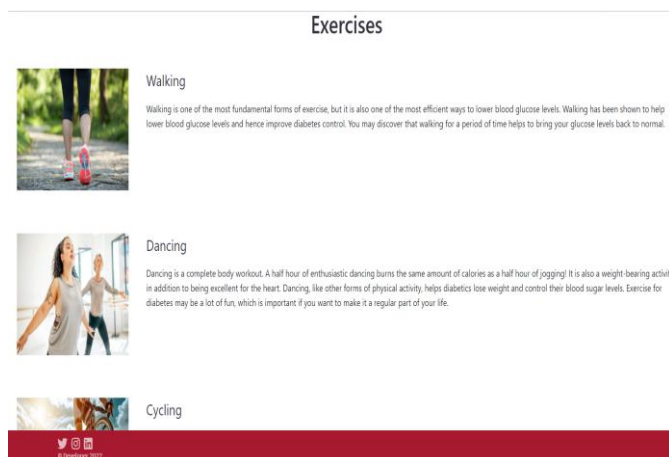


Figure -4: Exercises for pre-treatment

In addition to the above recommendations, glycemic index (which is a figure between 0-100 that denotes the ability of carbohydrates in food to increase the blood sugar level) of some common foods is mentioned for users to make informed decisions. Further, other general tips are presented to create more awareness as shown in Figure 5.

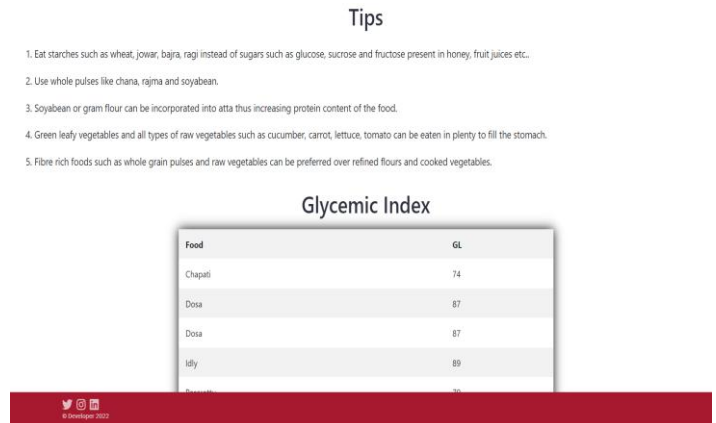


Figure -5: Other preventive measures

5. CONCLUSION

Diabetes is a critical and chronic disease that causes spike in blood sugar. Diabetic nephropathy, stroke, and failure of various organs, especially the kidneys, veins, and eyes can all be caused by undiagnosed diabetes. As a result, one of the most pressing medical issues in the world is the early detection of diabetes. The aim of the paper was to implement an end-to-end application that would detect diabetes early on and ensure pre-treatment through diet charts, exercises etc. Initially, the study performed comparative analysis on seven algorithms (Logistic Regression, Decision Tree, Random Forest, SVM, KNN, Adaboost and Gradient Boost) whose accuracy lied within the range of 67-80%. Thus, accuracy was improved to 82% by building an ensemble model of the five highest performing algorithms. In future: selecting a dataset bigger than the Pima Indians dataset, testing deep learning algorithms and improving feature extraction methods are various criterions to consider to obtain a better fitting model to improve accuracy.

REFERENCES

- [1] Z. Punthakee, R. Goldenberg, and P. Katz, "Definition, Classification and Diagnosis of Diabetes, Prediabetes and Metabolic Syndrome," *Can. J. Diabetes*, vol. 42, pp. S10-S15, 2018.
- [2] R. Vaishali, R. Sasikala, S. Ramasubbareddy, S. Remya, and S. Nalluri, "Genetic algorithm based feature selection and MOE Fuzzy classification algorithm on Pima Indians Diabetes dataset," in *Proc. International Conference on Computing Networking and Informatics*, Oct. 2017, pp. 1-5. M. Young, *The Technical Writer's Handbook*. Mill Valley, CA: University Science, 1989.
- [3] L. Lucaccioni and L. Iughetti, "Issues in Diagnosis and Treatment of Type 1 Diabetes Mellitus in Childhood," *J. Diabetes Mellit.*, vol. 06, no. 02, pp. 175-183, 2016.
- [4] "Type 2 Diabetes: a Review of Current Trends -," *Int. J. Curr. Res. Rev.*, vol. 7, no. 18, pp. 61-66, 2015

- [5] T. N. Joshi and P. P. M. Chawan, "Diabetes Prediction Using Machine Learning Techniques," *Ijera*, vol. 8, no. 1, pp. 9–13, 2018.
- [6] D. Sisodia and D. S. Sisodia, "Prediction of diabetes using classification algorithms," *Procedia Computer Science*, vol. 132, pp. 1578–1585, 2018.
- [7] S. Kumari, D. Kumar, and M. Mittal, "An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier," *International Journal of Cognitive Computing in Engineering*, vol. 2, pp. 40–46, 2021.
- [8] J. P. Kandhasamy and S. Balamurali, "Performance analysis of classifier models to predict diabetes mellitus," *Procedia Computer Science*, vol. 47, pp. 45–51, 2015.
- [9] S. Perveen, M. Shahbaz, A. Guergachi, and K. Keshavjee, "Performance analysis of data mining classification techniques to predict diabetes," *Procedia Computer Science*, vol. 82, pp. 115–121, 2016.
- [10] Vijayan V, Ravi K (2015) Prediction and diagnosis of diabetes mellitus—a machine learning approach, December, pp 122–127
- [11] Polat K, Güneş S, Arslan A (2008) A cascade learning system for classification of diabetes disease: generalized discriminant analysis and least square support vector machine. *Expert Syst Appl* 34(1):482–487
- [12] J. W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler, and R. S. Johannes, "Using the ADAP learning algorithm to forecast the onset of diabetes mellitus," in *Proc. Annual Symposium on Computer Application in Medical Care*, Nov. 1988, pp. 261-265.