**Diabetes Prediction in a Healthy Population Using Machine Learning**

Venkata Sneha Netuluri1, Venkata Varun Sai Nallapaneni1, Bindu Thokala1, Prathima Reddy Podduturi1

1School of Information Technology (Student), University of Cincinnati, Ohio, USA

[**netuluva@mail.uc.edu**](mailto:netuluva@mail.uc.edu)**,** [**nallapvi@mail.uc.edu**](mailto:nallapvi@mail.uc.edu)**,** [**thokalbu@mail.uc.edu**](mailto:thokalbu@mail.uc.edu)**,** [**poddutpr@mail.uc.edu**](mailto:poddutpr@mail.uc.edu)

**Abstract**

This research makes use of NHANES, the National Health and Nutrition Examination Survey, which generously supplied the raw information for this study. My GitHub repository contains both the dataset and the code for this project. Patients' chances of developing diabetes are predicted using this dataset's input parameters, which include things like age, glucose, blood pressure, insulin, body mass index, and so on. The patient's vitals are listed in each individual row of the data. All the patients here have a minimum age requirement of 21. The dataset includes 768 unique records for 9 feature sets. The following is a detailed description of each feature: Pregnancy shows the total number of pregnancies. Glucose displays the amount of glucose in the blood. Diastolic blood pressure is shown as blood pressure in mm Hg. The data represent the triceps skinfold's thickness in millimeters. The data specifies the amount of insulin in U/mL. BMI: The body mass index is shown in kg/m2. Diabetes Pedigree Function: This feature predicts the probability of diabetes based on family history. Age: This reveals the individual's age. The result reveals if the patient had diabetes or not (1 = yes, 0 = no).

**Introduction**

Diabetes is a metabolic disorder characterized by elevated blood glucose levels caused by either insufficient insulin production or insulin resistance. It's important to remember that diabetes is a complicated medical illness that may have serious outcomes. Globally, there are already more than 463 million people living with diabetes, and that number is expected to rise to 700 million by 2045, according to the International Diabetes Federation. Multiple complications, including cardiovascular disease, neuropathy, retinopathy, and renal failure, are associated with diabetes, making it a leading cause of death and disability across the globe. Having reached epidemic proportions, diabetes is a major public health concern across the world [1]. By spotting and treating the illness quickly, serious consequences may be avoided or at least mitigated. Since diabetes often has no noticeable symptoms in its early stages, identifying those at risk of acquiring the condition may be difficult. Therefore, it is crucial to create methods and technologies that can predict the chance of diabetes development in a healthy population.

Age, gender, genetics, lifestyle choices, and clinical features are only few of the variables that may affect diabetes development [2]. Obesity and inactivity can increase the danger. The creation of prediction models that consider a wide range of factors in determining an individual's diabetes risk has been made possible by recent advances in machine learning and data mining methods [3]. The models give the capacity to identify those at high risk of acquiring diabetes and provide individualized suggestions for preventing or intervening in the disease at an early stage. In recent years, the Finnish Diabetes Risk Score (FINDRISC) has become the most widely used model for making such predictions. (FINDRISC). To estimate one's 10-year risk of acquiring diabetes, the FINDRISC calculator makes use of non-invasive technologies [4]. To do this, we consider variables including age, BMI, waist circumference, exercise level, smoking status, and genetic predisposition to diabetes. The FINDRISC tool has been validated for use across different populations, and its efficacy in identifying those at high risk for acquiring diabetes has been established [5].

Several models have been developed to predict the probability of developing diabetes, including the Cambridge Diabetes Risk Score (CDRS), the Australian Type 2 Diabetes Risk Assessment Tool (AUSDRISK), and the Indian Diabetes Risk Score. (IDRS) [6]. The models in question incorporate supplementary variables, including but not limited to ethnicity, blood pressure, and fasting glucose levels, to enhance their precision.

Predictive models for determining the likelihood of developing diabetes have been created using machine learning methods including support vector machines (SVMs), decision trees, and other algorithms [7]. The models can process large datasets and deduce complex relationships between several risk variables, paving the way for the creation of credible risk assessments [8].

Wearable gadgets and other forms of mHealth technology have made it possible to track a person's activity level, sleep patterns, and food consumption in real time [9]. Including these innovations in prediction models may improve their accuracy and provide targeted advice for early intervention and preventative care.

Predicting the chance of diabetes occurrence in a non-diabetic population is, thus, an important metric in the early identification and mitigation of this complex metabolic condition. Methods from the fields of data mining and machine learning have been used to generate highly accurate prediction models for use in risk assessment. Multiple risk factors are accounted for in the models [10]. The models allow for the identification of those with a higher risk of acquiring diabetes, as well as the provision of individualized suggestions for preventing the disease and initiating care at an earlier stage. There is hope that the accuracy and utility of diabetes risk assessment models may be improved with the incorporation of mobile health technology and wearable devices into these models.

**Flow Model**

Data

Training dataset

Evaluation

Algorithm

Model

Testing dataset

Prediction

**Data collection**

In this article, we present a trove of NHANES statistics, where NHANES stands for the National Health and Nutrition Examination Survey. The NHANES is a research initiative that polls U.S. residents of all ages about their diet and overall health. Information on demography, habits, nutrition, and health problems like diabetes are all going to be included in our extensive data collection. To that end, we will be using this diabetes information in our research to determine how well we can forecast whether a given individual will develop diabetes.

**Data Preprocessing**

Our project's material is going to undergo the following preprocessing procedures: Data preparation is going to begin with a thorough cleaning of the raw data to get rid of redundancies, incorrect entries, and absent numbers. In this project we found large amounts of data from google, Facebook and other websites, so we selected a dataset for us from one of that source which is NHANES dataset and we have done preprocessing manually like we took small part of dataset feature columns like some categorical data like gender, numerical data like age and some text data. We replaced the blank values with null in some features and we removed some rows that contains inconsistent data and wrong values. We later split the data in to training and test data sets to train the machine and tested on the remaining part of data. We did scaling with standard scaler and tuned hyper parameters which we took are precision and recall for this and found the mean, standard deviation values for the parameters we have chosen or divided using code. We later found the accuracy score and misclassification rate by using the confusion matrix that is generated.

**Data Analysis**

Scikit learn is the data analysis library which we used in our project to analyze the data. Support vector machines, random forests, K nearest neighbors, Decision tree and gradient boosting technique are just some of the machine learning methods that can be implemented with the help of these tools. Separating information into test and training sets Data sets are usually segmented into training and testing sets so that the efficacy of a machine learning model can be evaluated. The model is going to be trained using the training set of data and tested on the testing set of data to assess how well it performed after training.

**Machine Learning Algorithms:** Decision trees, logistic regression, support vector machines, random forests, and gradient boosting technique are machine learning methods we used.

**Tuning Hyperparameters:** Hyperparameters are the factors that are defined before creating a machine learning model. The use of hyperparameter optimization is aimed at achieving the best possible results on the testing set by choosing the optimum collection of hyperparameters. Grid search is the method that we used for hyperparameter tuning here. Data analysis for a diabetes prediction model in a healthy population using machine learning typically involves the use of tools and software to implement statistical and machine learning algorithms, splitting the data into training and testing sets, and tuning hyperparameters to achieve optimal performance on the testing data set.

**Evaluation Metrics / Model Evaluation**

Common assessment criteria for a Machine Learning-based Diabetes Prediction model in a healthy population include the following:

Accuracy is defined as the fraction of test cases that were accurately labeled. Accuracy is the ratio of true positives (those that were both forecasted and confirmed to be positive) to the total number of positive predictions. Remember that this is the rate at which true positives occur relative to the total number of positives []. The F1-score is a superior metric to accuracy alone when there is a significant disparity between the groups because it considers both precision and memory. The typical discrepancy between observed and forecasted data is planned to be measured by our team using a statistic called the Root Mean Squared Error (RMSE) [10]. Metrics like accuracy, precision, recall, F1-score, RMSE, R2, and ROC-AUC are commonly going to be used when assessing the efficacy of a machine learning-based diabetes prediction model in a healthy population.

**Results**

**Dataset visualization**

Here we have added the main features of the dataset, where we can see the impact of every feature according to the color shade. With the use of a correlation matrix, we can see how each independent variable contributes to the whole. In this case, the attribute "glucose" is highly correlated with the predicted result. Apart from that, there is little to no association between any of thesis means it's conceivable such aspects won't be able to be ignored during model training.

These are the few dataset features which we used in our dataset to train and test the model, we considered pregnancy, glucose levels, blood pressure, skin thickness, insulin value, age, BMI and some more values based on their importance in finding the value whether the person can effect with diabetes or not.

**Chart, histogram

Description automatically generated**

**Fig : Features that are used in our dataset**

**Model’s Performance**

The table below shows the detection outcome of the selected model based on the adopted evaluation metrics. The decision tree obtained (Acc=0.7292), (Misclassification rate=0.2704), and gradient boosting obtained (Acc=0.75), (Misclassification rate=0.25). For the random forest algorithm, the accuracy rate obtained is 0.7447 and the misclassification rate is 0.25520. For the logistic regression, the accuracy rate obtained is 0.7239 and the misclassification rate is 0.2760. For the KNN, the accuracy rate obtained is 0.713541 and the misclassification rate is 0.28645. For the SVM algorithm, the accuracy rate obtained is 0.734375, and the misclassification rate is 0.265625. We found the most accurate results for the gradient boosting algorithm, so we considered it the main result for this project.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Performance Statistics   |  |  |  |  |  | | --- | --- | --- | --- | --- | | MODELS | Accuracy of the training set | Accuracy of the test set | Accuracy rate | Misclassification rate | | Decision tree | 0.852 | 0.729 | 0.7292 | 0.27084 | | Gradient boosting | 0.882 | 0.750 | 0.75 | 0.25 | | Random forest | 0.917 | 0.745 | 0.7447 | 0.25520 | | Logistic regression | 0.783 | 0.724 | 0.7239 | 0.2760 | | KNN | 0.79 | 0.71 | 0.713541 | 0.28645 | | SVM | 0.63 | 0.62 | 0.734375 | 0.265625 | |

Obtained values:

Accuracy Graph

Fig: Accuracy for all the selected models

Comparison with related studies:

|  |  |  |
| --- | --- | --- |
| S/N | Ref. | Accuracy (%) |
| 1. | C. D. Mathers and D. Loncar [1] | 0.55 |
| 2. | O. Tschritter, A. Fritsche [4] | 0.68 |
| 3. | J. P. Burke, K. Williams [8] | 0.70 |
| 4. | This study | 0.75 |

4.0 Conclusion and future works:

One problem that modern society has is that individuals don't eat a balanced diet that allows them to keep their blood sugar levels under control by limiting their consumption of sugary foods. Classifiers such as SVM, GBM, Random Forest, and Logistic Regression have all been used. In this study, we propose gradient boosting and compare it to other machine learning techniques using many data sets. With an average classification accuracy of 0.75, or 75%, the experimental findings demonstrate the efficacy of the gradient boosting approach. Thus, the suggested approach may be a great strategy to aid in the prevention of diabetes.

**References**

[1] C. D. Mathers and D. Loncar, “Projections of Global Mortality and Burden of Disease from 2002 to 2030,” PLoS Medicine, vol. 3, no. 11,p. e442, Nov. 2006.

[2] World Health Organization, “Definition and diagnosis of diabetes mellitus and intermediate hyperglycaemia: report of a WHO/IDF consultation,” 2006.

[3] American Diabetes Association, “Diagnosis and Classification of Diabetes Mellitus,” Diabetes Care, vol. 28, no. Supplement 1, pp. S37–S42, Jan. 2005.

[4] O. Tschritter, A. Fritsche, F. Shirkavand, F. Machicao, H. Haring, and M. Stumvoll, “Assessing the Shape of the Glucose Curve During an Oral Glucose Tolerance Test,”

Diabetes Care, vol. 26, no. 4, pp. 1026–1033,Apr. 2003.

[5] J. E. Shaw, P. Z. Zimmet, M. de Courten, G. K. Dowse, P. Chitson,H. Gareeboo, F. Hemraj, D. Fareed, J. Tuomilehto, and K. G. Alberti,“Impaired fasting glucose or impaired glucose tolerance. What best predicts future diabetes in Mauritius?” Diabetes Care, vol. 22, no. 3, pp. 399–402, Mar. 1999.

[6] N. Unwin, J. Shaw, P. Zimmet, and K. G. M. M. Alberti, “Impaired glucose tolerance and impaired fasting glycaemia: the current status on definition and intervention,” Diabetic Medicine, vol. 19, no. 9, pp. 708–723, Sep. 2002.

[7] M. A. Abdul-Ghani, K. Williams, R. A. DeFronzo, and M. Stern, “What Is the Best Predictor of Future Type 2 Diabetes?” Diabetes Care, vol. 30, no. 6, pp. 1544–1548, Jun. 2007.

[8] J. P. Burke, K. Williams, S. P. Gaskill, H. P. Hazuda, S. M. Haffner, and M. P. Stern, “Rapid Rise in the Incidence of Type 2 Diabetes From 1987 to 1996: Results From the San Antonio Heart Study,” Archives of Internal Medicine, vol. 159, no. 13, p. 1450, Jul. 1999.

[9] C. Lorenzo, K. Williams, K. J. Hunt, and S. M. Haffner, “Trend in the Prevalence of the Metabolic Syndrome and Its Impact on Cardiovascular Disease Incidence: The San Antonio Heart Study,” Diabetes Care, vol. 29, no. 3, pp. 625–630, Mar. 2006.

[10] S. M. Haffner, M. P. Stern, H. P. Haztjda, J. A. Pugh, and J. K. Patterson, “Hyperinsulinemia in a Population at High Risk for Non-Insulin Dependent Diabetes

Mellitus,” New England Journal of Medicine, vol.315, no. 4, pp. 220–224, Jul. 1986.