Development of Efficient Machine Learning Model to Detect Muliebrous

Rishabh Nagar, Rinisha Jain, Amogh Shukla

Vellore Institute of Technology, Vellore

Corresponding author: Rishabh Nagar, Email: rishabhbhavesh99r@gmail.com

The Diabetes predicting machine learning models developed so far Diabetes risk calculators have a strong negative predictive value, which means they can help identify people who are unlikely to develop diabetes. Tests that reflect beta cell function may assist determine the type of diabetes or the requirement for insulin indefinitely. Machine Learning models can help predict whether or not a female is suffering from diabetes while assessing different health conditions such as; increased glucose level, age, family hereditary and blood pressure. Our project focuses on making machine learning models which can predict if a person suffers from diabetes while using specific health conditions as constraints. So different machine learning models will be used in order to get concluding results and whichever model serves to be most accurate will be used to improve the prediction system of diabetes amongst female patients. The project uses several concepts of machine learning, and the data used are databases obtained from national authorities to test and verify the results. Diabetes, if ignored, can cause various life threatening situations for the people suffering from it and hence we aim to help predict diabetes with accuracy among the female population.

Keywords: Diabetes, Machine Learning, Medical Visualization.

2023. In Saroj Hiranwal & Garima Mathur (eds.), *Artificial Intelligence and Communication Technologies*, 895–903. Computing & Intelligent Systems, SCRS, India. https://doi.org/10.52458/978-81-955020-5-9-85

1. Introduction

The Diabetes predicting system for females is a software application that will help predict whether or not a female is suffering from diabetes while assessing different health conditions such as; increased glucose level, age, family hereditary and blood pressure. Our project focuses on making machine learning models which can predict if a person suffers from diabetes while using specific health conditions as constraints. So different machine learning models will be used in order to get concluding results and whichever model serves to be most accurate will be used to improve the prediction system of diabetes amongst female patients. The research uses several concepts of machine learning, and the data used are databases obtained from national authorities to test and verify the results. We will be taking these factors as constraints. These constraints are:

- Pregnancies
- Glucose
- Blood Pressure
- Skin Thickness
- Body Insulin
- Body Mass Index
- Age
- Likelihood of diabetes

By studying these constraints, we can predict the outcome as “Yes: Suffers from diabetes” or “No: Does not suffer from diabetes”. These data will be taken from an online database. We will be using different machine learning algorithms and compare each one so that we can provide with the best outcome. Also, since pregnancy is also a constraint, the models only work for females.

1.1. Objective

The goal is to use certain standard medical measurements included in the dataset to determine whether a patient is diagnosed with diabetes or not. Utilising Machine learning concepts to create a diabetes prediction system Applying concepts of software engineering to bring out efficient execution of the desired project Coming up with the best suitable algorithm to get accurate prediction Help doctors speed up the diagnosis process of diabetes among female patients.

1.2. Motivation

Diabetes is one of the major diseases that many of the adults in India suffer from. It is a metabolic disorder which is identified by the high blood sugar level. There are many factors that lead to this disease. Some of them being: increased glucose level, age, family hereditary and blood pressure. By studying these factors, one can successfully predict whether a person will suffer from Diabetes or not. Our project focuses on making machine learning models which can predict if a person suffers from diabetes. Diabetes, if ignored, can cause various life- threatening situations for the people suffering from it and hence we aim to help predict diabetes with accuracy among the female population.

2. Literature review

[1] is an illness prediction method that focuses on predicting by evaluating the user’s condition diagnostics that the user offers as input. The section analyzes the user’s symptoms as input and outputs a probability of illness as an output. The Naive Bayes system is frequently used to forecast sickness. It use linear regression and decision trees and predict metabolic illnesses, Malaria, Jaundice, West Nile virus,

and Tuberculosis. This method use ml methods to accurately anticipate long-term illness outbreaks in illnesses areas. Researchers used actual hospital data from central China to test the improved prediction algorithms. Deep convolutional neural networks oriented for risk forcasting in multimodal subsystems (CNN-MDRP) method is proposed as a new holistic sickness risk prediction model.

In [2,] the most inspiring areas of study that have grown most popular in medical associations are data mining and machine learning. It also plays an important role in investigating potential patterns there in medicinal science and services association, which benefits all parties involved in this field. Using datasets such as categorization in the health sector, this study aims to build a diagnosis model of prevalent chronic diseases based on symptoms. We will use algorithms such as Random Forest and Naive Bayes in this project, that could be used for health care diagnosis. Against determine which classifier has the highest accuracy, the performance of the classifiers are compared to one another.

Chronic kidney disease (CKD), also known as chronic renal disease, is a condition that affects the kidneys. Chronic renal disease [3] is a group of illnesses that affect your kidneys and reduce their ability to keep you healthy. High blood pressure, anaemia (low blood count), weak bones, poor nutritional status, and nerve damage are all possible problems. Early detection and treatment can typically prevent the progression of chronic renal disease. The term "data mining" refers to the process of extracting knowledge from big databases. The goal of data mining is to analyse previous data to find regular patterns and better future judgments. It is the result of the convergence of numerous current trends: the decreasing cost of massive data storage devices and the rising ease with which data may be collected through networks; the development of more reliable and efficient machine learning algorithms to handle this data; and the decreasing cost of processing power, allowing the use of computationally expensive data analysis approaches.

In the medical evaluation and prognosis process, data mining algorithms have been extensively used. These data mining algorithms have been used to analyse a large number of medical data

[4]. Overweight and obesity, as well as an unhealthy lifestyle, eventually reflect the threat and occurrence of liver-related disorders in general. In this study, patient records sets are examined for the predictive model of a person having a liver ailment using only a widely studied classification algorithm. Given that several systems are in place to evaluate patient and classification algorithm data, the more essential element here is to anticipate the same outcome with an elevated rate of accuracy.

Lauren N Carroll and Alan Au's Comprehensive Review Journal of Biomedical Sciences, April 2014. The goals of this paper [5] were to: (1) find public health users' needs and requirements for communicable diseases information visualisation tools; (2) identify existing communicable diseases information visualisation tools and characterise their structure and characteristics; (3) identify similarities among approaches utilized in various types of data; and (4) describe tool design and evaluation effort in order and obstacles to tool adoption. The structure of the tools was irregularly documented, and instead just a few of the tools in the review included usability studies or dissemination plans.

Blip is a new technology platform that integrates and exhibits diabetes information from different devices in a regular, user-friendly manner. The goal of this research [6] was to see how effectively people and caretakers of diabetic children could use this app. For three months, individuals (n = 35) and caregivers of children with Type 1 diabetes (n = 30) who were using an insulin pump for more than a year were provided access to the system. At the outset, 97 percent of those surveyed agreed that individuals should be capable of comprehending glucose data. Despite their value for shared responsibility, 43 percent of participants never retrieved pump data at baseline, and only 9.

A predictive algorithm aims to stratify patients based on the probability of achieving the targeted outcome. The model[7] then makes it possible to identify patients who are more likely to encounter an event, which might also lead to treatment adjustments for the particular patient. The prediction model's outcome variable can be anything, such as the risk of developing a side effect, the probability of surviving till a certain time point, or the probability of a tumor progression. We can classify outcome variables

either as continuous or categorical. Continuous variables are numerical values that can be anticipated using regression models. Binary classification is used when outcome has two categories, and typical techniques involve decision trees and logistic regression.

Heart disease also refers to conditions defined by narrowed or blocked blood vessels that can result in a cardiac arrest, chest pain, or pulmonary embolism. This study[8] presents a machine learning algorithm-based model for detecting heart disease. The Agile Model was used in this research, which includes parallel planning, requirements analysis, developing, programming, screening, and paperwork throughout the process of production. In this paper, four machine learning algorithms were used to train a Heart Dataset. This model was implemented to the online platform using flask (a Python framework), and it makes a prediction based on 13 user inputs. The model is constructed with the Python language and flask. This paper employs a Decision Tree Classifier Algorithm, and the prediction results indicate an accuracy of approximately 68.83 percent.

Empowered consumers in deregulated markets [9] expect innovative and customized services. As an outcome, there is a great need for enterprises to develop innovative models to improve consumer satisfaction and gain a competitive advantage over others. Using data mining and statistical techniques, we will be able to develop a number of novel modeling techniques. Predictive analytics and predictive analysis can improve customer relationship building. In this article, we discussed the use of data mining in business analytics, as well as the use of data mining languages like python and R in the development of forecasting analytics.

This study [10] explores how to assist diabetics in reflecting on and trying to improve their own health practices by collecting and imagining health-related information. They introduced diabetics to a new type of data collection, photography, to replenish the data they usually collect, sugar levels. Diabetics take pictures of their meals, exercise, work, play, and anything else that they believe has an effect on their health. This creates integrated data visualizations by combining quantitative glucose metrics with qualitative portraits of action. They hope that to do so, they can make the relationship between physiology and behavior a topic for discussion and reflection. More importantly, they hope that diabetics who have viewed these data will begin to develop new interpretations of their lifestyles, which will eventually lead to healthier activities.

3. Problem Statement

Diabetes is among the leading causes of chronic illness in people all around the world, particularly in women. It is difficult for medical practitioners to forecast in females, especially before pregnancy, because it is a complex task that necessitates competence and a greater level of understanding. An automated medical diagnosis system would improve medical efficiency while simultaneously lowering expenses. We'll create a system that can quickly learn the rules for predicting a patient's risk level based on the information provided about their health. The goal is to use data mining techniques to identify hidden patterns that are relevant to heart disorders and to forecast the presence of heart disease in patients whose presence is valued on a scale. Diabetes testing prediction necessitates a large amount of data that is too complex and enormous to collect and interpret using traditional methods. Our goal is to develop the most appropriate machine learning technique for predicting diabetes illness in females that is both computationally efficient and accurate. To extract hidden patterns and relationships from big databases, data mining combines statistical analysis, machine learning, and database technology.

4. Proposed Method

Users enter information like, pregnancy, glucose, blood pressure, insulin, skin thickness, BMI, diabetes history and age, along with the option to choose the classification model. The algorithm then uses the training data that was given to it during initialization, and predicts possible graphs for the information entered by the patients and compares them with the data used in the initialization process. This allows the patient to compare their information with the standard data that is derived from online websites. Using this information, the likelihood of diabetes can be calculated and predictions can be made. Along with this an option to give feedback is available, in order to improve on the prediction method and on the 5 overall prediction. This all allows the prediction algorithm to learn and improve on the accuracy of the model.

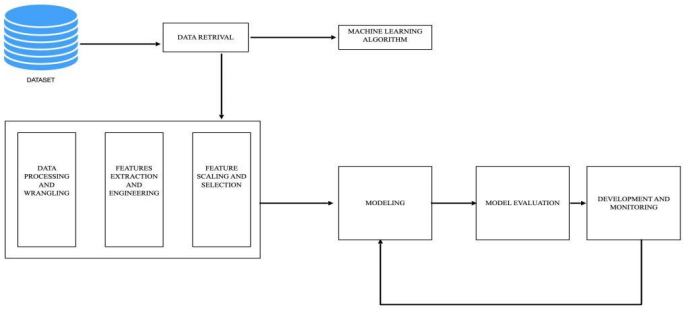

Fig 1: Flowchart of Proposed Work

5. Methods

5.1. Importing the Data set and Libraries

To begin, we will import all of the libraries required for diabetes prediction. After that, we'll use the pandas library to import our data set. We shall now conduct exploratory data analysis (EDA). It's a method of displaying, summarizing, and evaluating data that's concealed behind rows and columns.
Because the data contains numerous zeros and the values of glucose, blood pressure, skin thickness, insulin, and other attributes cannot be zero, the zeros are converted to mean and median values for the characteristics. The modified data set will then be exported. The Machine Learning Models are trained using the new Data set.

5.2. Data Visualization using the Different Plots

A pair plot is a graph that depicts the representation of paired relationships in a data set. The pair plot function creates a grid of Axes with a single row and single column for each variable in the data on the y-axis. The relationship between the different attributes is then described with a heat map of all the attributes.

5.3. Splitting Training and Testing Data set

To begin, the data set is divided into two sections: train (614[0.8]) and test (154[0.2]), after which the model is trained on train section and tested on the test section. The model predicts the outcome of test data once it has been trained. Then the Confusion Matrix and the Model's Accuracy, Precision, and Recall is calculated.

5.4. Validation and Comparison Of Different Machine Learning Algorithms

On comparing all machine learning techniques' accuracy, precision, and recall. The diabetic data set was best fitted by logistic regression. It has a precision of roughly 95.6% percent, which can be improved by using the K-fold Validation Method.

6. Validation

The hold-out method, as well as the k-fold cross validation procedure, were used in a number of research to test the model's functionality (Kohavi, 1995; Bengio and Grandvalet, 2005; 9 Kim, 2009; Chen et al., 2016; Refaeilzadeh et al., 2016; Yang et al., 2016, 2018; Su et al., 2018; Tang H. et al., 2018). Depending on the goal of each case and the complexity of the data, we can use several approaches to address the problem. The train set and the test set are separated in the hold-out method. The model is evaluated using the testing set, and the machine learning model is trained using the training set (Kim, 2009). This methodology was used to see if such methods are universally applicable in this study.

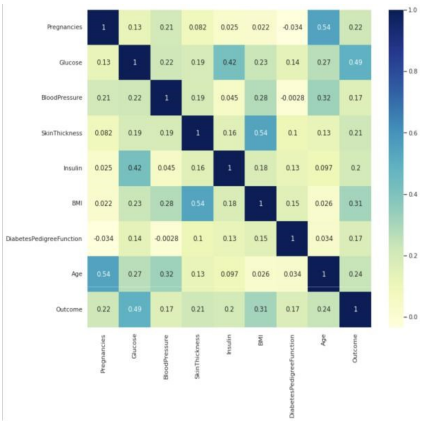

Fig 2: The Heatmap for All The Attributes

6.1. K-Fold validation method

K-fold cross validation method: In k-fold training and validation strategy, each full dataset was used to train the classifier the classifier (Kim, 2009). The dataset would be segmented across k portions, such as folds, first. During the training process, the method includes k-1 flips to fit the class and onefold to confirm it. Evey fold will constitute as a test set for all of this operation, which would be executed k times. The average of all the folds' medical reports is the final result. Regression Model is clearly the most efficient measurement.

7. Result

The introduction of two of the most efficient and predictive approaches, Random Forest and Logistic Regression, to help individuals at high risk of developing Diabetes Mellitus, was a significant contribution of our research. Both traditional statistical models and modern learning-machine techniques are implemented in this study. The problem of unbalanced data was resolved using the altered technique as well as the class weight method. The models were able to identify persons with Diabetes Mellitus with high precision and specificity.

Table 1: The comparision table of different models

	Decision	CNN	Logic Regression	Naive Bayes	Random Forest	SVM
Accuracy	63.43534559	77.51339035	81.34234033	78.43290340	77.32490340	89.51334939
Precision	81.2540439	84.54490423	83.45230934	83.5354903	82.3429034	82.34290493
Recall	72.453493	83.4250	80.34324933	85.34343094	80.43254945	90.34243344

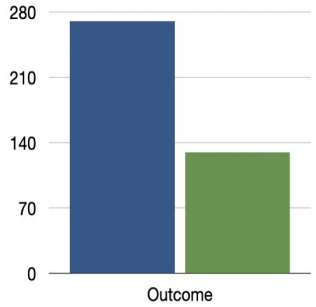

Fig 3: Count Plot of Outcome

To highlight the risk characteristics, these estimation techniques are created and tested on a random group of patients with Diabetes Mellitus. These models are used as software to help professionals analyze a patient's likelihood of developing Diabetes Mellitus in females. Every year, diabetes kills more people of 9000 people all around the world. As a result, recognizing diabetes early is critical for 's overall.

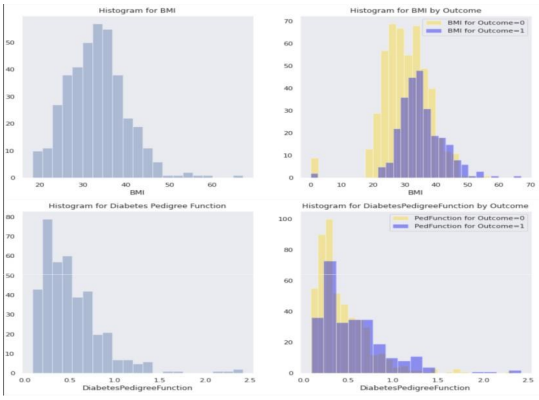

Fig 4: Histograms of BMI Diabetes Pedigree Function

This study employed a machine learning algorithm to predict diabetes. This study uses machine - learning techniques to monitor diabetes affected by a range of medical data. According to my data, two-fold cross validation demonstrated effectiveness from 95.5 percentage to 99 percent.After investigating this pattern, we observed that by employing five internal validations, the accuracy may be enhanced to 99.5 percent, which is much superior than other diabetic mellitus forecasting techniques now and in use. Both medical personnel and common citizens will profit out from proposed system.

A count plot is a histogram across a category variable rather than a quantitative variable. Here is the count plot for the Outcome attribute, which specifies that if 0 is obtained, the person does not have diabetes, and if 1 is obtained, the person does.

A Heatmap is a representation of data in the form of a map or diagram in which data values are represented as colors. Here, our Heatmap shows the comparison of all the attributes of the dataset used along with their color variation.

8. Discussion

8.1. Conclusion

Diabetes (that must be avoided before it causes individuals suffering) kills millions each year and across the world. As a consequence, accurate diagnosis of diabetes is useful for successful treatments. A deep neural network is shown in this study to determine diabetes. The deep neural network technique was included in this study to determine diabetes depending on a number of medical factors. For two-fold cross-validation, we observed that accuracy increased from 68 percent to 94 percent. As a result of examining this pattern, we have concluded that by using five cross validation, the accuracy may be increased to 98 percent, which is significantly greater than other methods currently used to forecast diabetes mellitus.

8.2. Future Work

Using our dataset, we have used conventional statistical models and contemporary learning methods. We have used adjusted-threshold strategy and the classed weight method to address the problem of outlier data. Our algorithms have a high sensitivity and a strong ability to detect people with Diabetes Mellitus. However newer hybrid techniques have been discovered and can be used to predict the disease in an even better way. These techniques may be used to validate these results. These forecasts are created and tested on a random group of patients with Diabetes Mellitus to represent the risk patterns. These models can be set up in an online computer software to assist doctors in determining a patient's risk of acquiring Diabetes Mellitus in females.

References

[1] Kedar Pingale, Sushant Surwase, Vaibhav Kulkarni, Saurabh Sarage, Prof Abhijeet Karve (2019). Disease Prediction using Machine Learning International Research Journal of Engineering and Technology, Volume: 06 Issue: 12 | Dec 2019.

[2] Ashish Kailash Pal, Pritam Rawal, Rahil Ruwala, Prof. Vaibhavi Patel (2019). Generic Disease Prediction using Symptoms with Supervised Machine Learning. International Journal of Scientific Research in Computer Science, Volume 5 Issue 2 | March-April 2019.

[3] Parul Sinha, Poonam Sinha (2015). Comparative Study of Chronic Kidney Disease Prediction using KNN and SVM. International Journal of Engineering Research and Technology (IJERT), Vol. 4 Issue 12, December- 2015.

[4] Durai Vasan (2019). Liver disease prediction using machine learning. International Journal of Advance Research, Ideas and Innovations in Technology, Volume 5 Issue 2 | March-April 2019.

[5] Ms. Priti V. Wadal, Dr. S. R. Gupta (2014).
Predictive Data Mining For Medical Diagnosis: An Overview Of Heart Disease Prediction.
International Conference on Industrial Automation and Computing, (ICIAC-12- 13th April 2014).

[6] Jenise C. Wong, MD, PhD, Aaron B. Neinstein, MD, Howard Look (2017). Pilot Study of a Novel Application for Data Visualization in Type 1 Diabetes. Journal of Diabetes Science and Technology.

[7] FrankJ.W.M.Dankers, Alberto- Traverso, LeonardWee, and SanderM.J.van Kuijk (2019). Pre-diction Modeling Methodology. study. Fundamentals of Clinical Data Science.

[8] O.E. Taylor, P. S. Ezekiel, F.B. Deedam-Okuchaba (2019). A Model to Detect Heart Disease using Machine Learning Algorithm IJCSE International Journal of Computer Sciences and Engineering, Volume-7, Issue-11 | Nov 2019.

[9] Dr.V.V.Narendra Kumar, Dr.K.Kondaia , Regula Thirupathi (2017). Building predictive model by using python and r for enhancing consumer satisfaction and advantages International Journal of Advanced Technology in Engineering and Sciences,Volume 5 Issue 3 | March 2017.

[10] Jeana H Frost and Brian K Smith (2012). Visualizing health practice to treat diabetes Extended abstracts of the 2002 Conference on Human Factors in Computing Systems, CHI 2002, Minneapolis, Minnesota, USA,, April 20-25, 2002.