Comparative Analysis of Predicting the Diabetic Disease Using Machine Learning Techniques

J Revathy\textsuperscript{a,1} and D Selvanayagi\textsuperscript{b}

\textsuperscript{a}Research scholar, Department of Computer Applications
\textsuperscript{b}Asst. Prof, Department of Computer Applications
\textsuperscript{a, b}Vellalar College for women, Erode, Tamil Nadu, India

Abstract. Machine Learning is concerned with the making of calculations and methods that use PCs to learn and acquire insight, using the related knowledge available. This work is focused on machine learning approaches for predicting diabetic disorders, using datasets from Predict the Diabetic Diseases. A web-based comparative analysis of multiple machine learning algorithms (Decision Tree, Support Vector Machine, k-Nearest Neighbor, and Logistic Regression) is utilized in this paper, to assess their performances in recognizing reliable models for detecting diabetic disease. To see the effects of adding more features to the classification model, three performance measures were chosen: F1-Measure, Precision, and Accuracy.

Keywords: Machine Learning, Decision Tree, Support Vector Machine, Logistic Regression, Diabetes

1. Introduction

Diabetes diagnosis has been carried out utilizing multiple machine learning algorithms to forecast the illness of a patient and to improve treatment such as Decision Tree (DT), Support Vector Machine (SVM), k-Nearest Neighbor (k-NN), and Logistic Regression. Diabetes Mellitus (DM) is an ongoing condition described by the powerlessness of the body to use glucose. Early discovery of the illness brings down clinical usage and the probability of individual’s genuine medical issues. DM research is a critical methodology for separating the needed information from the gigantic database of diabetes-related information. The financial cost is one of the main concerns in clinical science study, and it normally creates enormous volumes of information.

Types of Diabetes

Type 1 Diabetes - The body doesn't fabricate sufficient insulin in this kind of diabetes. Insulin-subordinate diabetes, adolescent diabetes, and beginning stage diabetes are the terms used to depict this sort of diabetes. Type 1 diabetes, for the mostly part,
occurs before the age of 40, for example in early adulthood or during puberty, and the patient requires insulin shots for life.

Type 2 Diabetes - The body doesn’t make sufficient insulin or the cells in the body are insulin safe. A few people with type 2 might have the option to control their indications by reducing body weight, following a decent eating regimen, and intently checking their blood glucose levels. As Type 2 diabetes deteriorates after some time, the patient will very likely need to take insulin, for the most part in tablet structure. Obesity, absence of physical work, and eating unhealthy food lead to type 2 diabetes.

Type 3 Diabetes (Gestational Diabetes) – Women develop this during pregnancy. A few ladies’ blood glucose levels rise since their body doesn’t produce sufficient insulin to move all of the glucose into their cells, bringing about consistently rising glucose levels. Most of ladies, with gestational diabetes, might deal with their diabetes with physical work and a sound eating routine. About 10% and 20% of them may require blood glucose controlling medicine.

The focus of this study is to review various diabetes detection mechanisms on the Type 3 diabetes to identify this, Naive Bayes (NB), SVM, and DT classification methods are employed and assessed on Pima Indian Diabetic dataset (PIDD). Using several criteria, the experimental performances of these algorithms are compared, and they all attain good accuracy [1]. This study examines the familiar metrics such as F1-score, Precision, Recall and the Accuracy rate of various diabetes detection mechanisms [2].

2. Related Work

The research work [3] has collected the diabetic dataset, and various machine learning techniques are implemented. The Logistic Regression provides the highest accuracy of 96% and AdaBoost is found to be the best model with an accuracy rate of 98.8%. With two separate datasets, the accuracies of machine learning algorithms are compared.

Calisiretal [4] surveyed diabetic illnesses and diagnosed them using the Linear Discriminant Analysis (LDA) and Morlet Wavelet Support Vector Machine (MWSVM) classifiers. The LDA – MWSVM automatic diagnosis method for diabetes achieves extremely promising results in classifying possible diabetic individuals. The employed LDA – MWSVM automatic diagnosis system is thought to be particularly useful to clinicians, in making final decisions about their patients.

Georga et al [5] applied its capability to the table for smooth, worldwide, and meager answers for nonlinear regression problems. The Support Vector Regression (SVR) approach is pushed for multivariate prescient investigation of glucose digestion. They estimated the impact of each contribution on forecast exactness, utilizing different information circumstances, and utilizing the present moment and the long haul expectations. Finally, in addition to predicting values, the challenges of anticipating hypo/hyperglycemia episodes are also warned during experimentation.

Geetha. G et al [6] have utilized Random Forest (RF) for the prediction of diabetics. After separating and analyzing the training and testing data, this method provides us an approximate result. When compared to Naive Bayes, this approach has a substantially higher efficiency. As a result, the suggested method is highly effective for both diabetic analysis and prediction.

Deberneh et al [7] had indicated Type 2 diabetes occurrence prediction model that predicts as normal, pre-diabetes, or diabetes in the following year (Year+1). The
forecast models were made utilizing LR, RF, XGBoost, SVM, and group classifiers (CIM, ST, and SV), FPG, HbA1c, fatty substances, BMI, gamma-GTP, orientation, age, uric corrosive, smoking, drinking, actual work, and family ancestry were among the elements picked. The model gives critical data on the event of type 2 to the two specialists and patients early.

Kavakiotis et al [8] have surveyed many machine learning and data mining approaches used in DM. This work includes clinical and biological information that leads to more in-depth exploration of diagnosis. Polat et al [9] have depicted the impacts of master frameworks and AI methods for the analysis of diabetes. Settling on choices in the clinical business can be troublesome at times now and then. Clinical dynamic order frameworks permit clinical information to be investigated significantly quicker and in more detail. An interesting cross breed approach in the light of PCA and ANFIS is proposed in this paper for the determination of diabetes disease. The goal of diagnosing diabetics was applied to ANFIS with PCA in the exploration revealed in this work, and the most reliable learning approaches were assessed. On the diabetes sickness dataset, tests were embraced to analyze diabetes illness naturally, utilizing ANFIS and PCA.

Alghamdi et al [10] have applied AI techniques for anticipating occurrence diabetes, utilizing cardiopulmonary wellness information. Among the five characterization models utilized, the Random Under-Sampling procedure yielded improvement. The SMOTE approach, then again, exhibited an impressive improvement in the forecast capacity of all characterization models, as the rates used were step by step expanded. In every one of the models, the Random Forest and NB Tree models beat the others (Kappa, Recall, Precision and Specificity).

Faruque et al [11] have examined the early discovery of diabetes by thinking about a few gamble factors related with the disease. Foreseeing diabetic onset can be pretty much as straightforward as removing information from a genuine medical services dataset. To estimate diabetes mellitus, tests were directed on grown-up populace, utilizing four normal AI calculations SVM, NB, k-NN, and C4.5. The exhibition of the C4.5 is a lot better than the other AI calculations, for the arrangement of diabetic information. Soni et al [12] have applied machine learning algorithm to design and implement diabetes prediction such as SVM, k-NN, RF, DT, LR, and Gradient Boosting classifiers. The exploratory outcomes can assist the patient’s well-being, with caring experts settling on early forecasts and choices, to fix diabetes and save individuals lives.

Kadhm et al [13] have proposed a fast and accurate diabetes prediction algorithm. For each PID dataset, the suggested system used 768 instances with 8 attributes. Furthermore, by partitioning the dataset into subsets, an optimal classification result was achieved. The suggested approach concentrated on the analysis and classification of characteristics. The results, of the studies, demonstrated the benefits of using the suggested system's algorithms, by achieving a greater categorization rate than competing systems.

Tigga et al [14] have applied six different AI, and the outcomes were thought about. Tests were run on information gained on the web, and disconnected surveys, that included 18 diabetes-related inquiries. The exactness of RF is 94.10%, which is the best among the others. For the PIMA dataset, RF likewise, gives the greatest exactness. Each of the models showed great outcomes for greater accuracy, review responsiveness, etc.
Nnamoko et al. [15] have collected a variety of factors that influence their potential to improve performance, particularly at the base level. The basis classifiers are chosen from five different machine learning algorithm families. As a result, each classifier would generate models based on its operational features. Despite the fact that none of them produced any major improvements at the base level, the sum of their unique biases contributed to a greater understanding of the classification problem at the ensemble level, ultimately leading to significant improvement. If such pre-processing experiments had been carried out, their outcomes might have been different. However, the findings of this study show that advances were achieved as a result of feature selection applied to heterogeneous base learners, rather than data pre-processing.

Rajeswari et al. [16] has splitted the mathematical properties (hazard factors) of the PID informational collection, the recommended strategy - Fuzzy rationale based Associative Classification to resolve the issue of limit esteem confusion. It doesn't require proficient parceling or reach fixing information. This is reachable in the light of the fact that the MFs utilize normal limit parting to break the base and the most extreme worth of the characteristics into different language phrases. Whenever the idea of the FARS, created by the recommended strategy is contrasted with the exception pruning measurements - Lift, Leverage, and P-Value, obviously the proposed technique delivers a decent model. The proposed procedure is superior to the fresh strategy in deciding the specific gamble factors like Age, Glucose, DPF, BMI, and BP, as well as their perilous qualities, to foresee pre-diabetes.

Agnal and Saraswathi [17] have applied Naive Bayes Classifier's methodological approach to reliably and precisely diagnose diabetes. The proposed approach outperforms Hadoop Technology, in handling both large and small volumes of data. As a classification algorithm, this procedure is quick, secure and easy to change, and produces accurate results. By improving accuracy, it can be beneficial in medical research.

Vijiya Kumar et al. [18] have introduced determination of a hereditary anomaly at the beginning phase, as one of the necessary genuine clinical worries. Irregular Forecast calculations have been analyzed and evaluated on an assortment of scales during this undertaking. The objective of this venture is to make a framework that can precisely gauge diabetes in a patient early, using an AI strategy that gives advance assistance to anticipating diabetes exactness rates. The consequences of different AI calculations are displayed in Table 1, for the PIMA Indian Diabetic Datasets.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>F1-Score</th>
<th>Precision</th>
<th>Accuracy</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear regression</td>
<td>0.722732</td>
<td>0.754323</td>
<td>0.843782</td>
<td>0.693683</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>2.313075</td>
<td>0.771025</td>
<td>0.884578</td>
<td>0.761037</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>2.648536</td>
<td>0.672584</td>
<td>0.703565</td>
<td>0.662134</td>
</tr>
<tr>
<td>SVM</td>
<td>2.627466</td>
<td>0.875822</td>
<td>0.917324</td>
<td>0.846212</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>3.164824</td>
<td>0.800103</td>
<td>0.815395</td>
<td>0.791206</td>
</tr>
<tr>
<td>k-NN</td>
<td>1.129604</td>
<td>0.274685</td>
<td>0.318189</td>
<td>0.282401</td>
</tr>
<tr>
<td>Random Forest</td>
<td>1.075652</td>
<td>0.265264</td>
<td>0.282280</td>
<td>0.268913</td>
</tr>
</tbody>
</table>
3. Conclusion

The first point of this advancement was to propose and achieve Diabetes Prediction Using Machine learning strategies and investigation of a strategy has been classified. The proposed methodology approach uses different categorization such as Linear regression, Decision Tree, Logistic Regression, SVM, Naive Bayes, KNN, random Forest classifiers are used. The comparative analysis and results illustrate the early way to cure diabetes on earlier prediction to save a human life with machine learning Techniques.

![Machine Learning Diagram](image)

**Figure 1.** Various Machine Learning Algorithm used for diabetes prediction.

References


