Comparative Analysis of Machine Learning Models for Diabetes Mellitus Type 2 Prediction

Leila Ismail, Member, IEEE, and Huned Materwala
Distributed Computing and Systems Research Laboratory
Department of Computer Science and Software Engineering, College of Information Technology, United Arab Emirates University, Al- Ain, Abu-Dhabi, 15551, United Arab Emirates
Corresponding author: Leila Ismail (e-mail: leila@uaeu.ac.ae)

Abstract—Diabetes is one of the top 10 causes of death worldwide. Health professionals are aiming for machine learning models to support the diagnosis of diabetes for better healthcare and to put in place an effective prevention plan. In this paper, we conduct a comparative analysis of the most used machine learning models in the literature to predict the prevalence of diabetes mellitus type 2. We evaluate the models in terms of accuracy, F-measure and execution time with and without feature selection using a real-life diabetes dataset. The detailed analysis is in the paper.

Keywords—artificial intelligence, classification models, diabetes mellitus type 2, health informatics, machine learning models

I. INTRODUCTION

Diabetes Mellitus (commonly referred to as Diabetes) is one of the top 10 causes of death worldwide with 4.2 million deaths in 2019 [1]. According to the International Diabetes Federation (IDF), the total number of adults (20-79 years) with diabetes will increase from 463 million in 2019 to 700 million in 2045, an increase of 51.18% [1]. In 2019, USD 760 billion was spent on diabetes health expenditure, which is 10% of the total spending on adults. There are three main types of diabetes: type 1, type 2, and gestational diabetes [2]. The number of people with type 2 diabetes is increasing compared to ones with type 1 and gestational diabetes [1]. In 2019, more than 1.1 million children and adolescents were suffering from type 1 diabetes, while 374 million people were at increased risk of type 2 diabetes. Those statistics related to type 2 diabetes are frightening. Therefore, it becomes crucial to address this issue and to provide health professionals with a support system to predict the prevalence of diabetes in patients for effective healthcare. In addition, the system would help to put in place an efficient preventive plan at the national and global level.

Type 2 diabetes is thought to prevail in an individual from an interaction between lifestyle, psychosocial, medical conditions, and genetic risk factors [3, 4]. If type 2 diabetes is not diagnosed at an early stage, it can lead to life-threatening situations such as neuropathy, heart and blood vessel disease, kidney damage, eye damage, hearing impairment, skin conditions like bacterial and fungal infections, and Alzheimer [5]. To date, there is no permanent cure for diabetes, and the patients have to rely on healthy lifestyle adaption and timely medication. Machine learning classification models have shown to be a promising approach to support health professionals for the prediction of diabetes. A classification model predicts the class of a given input data [6]. We develop machine learning classification models using a labeled training dataset, and we evaluate them using a testing dataset. A dataset is said to be labeled if both input and output variables are used for training. The input variables are the features of the dataset, while the output variables are the class labels. In the case of diabetes, features are the risk factors, and the class labels are diabetic and non-diabetic.

The work in the literature on diabetes prediction evaluates the performance of classification models using heterogeneous datasets and evaluation metrics. In this paper, we conduct a comparative analysis of the most used classification models in the literature, Decision Tree (DT) [7], Naive Bayes (NB) [7], k-Nearest Neighbor (k-NN) [8], Support Vector Machines (SVM) [9] and Artificial Neural Network (ANN) [10]. This is in a unified setup using the University of California Irvine (UCI) diabetes dataset [11] that consists of 12 features and 65,840 observations. We evaluate the performance of the models in terms of accuracy, F-measure and execution time with and without feature selection.

The rest of the paper is organized as follows. In Section 2, we present an overview of the related work. Section 3 explains the classification models used in this study. The experimental setup, experiments, and the results’ analysis in terms of accuracy, F-measure and execution time are described in Section 4. Section 5 concludes the paper.

II. RELATED WORK

Several research efforts in the literature evaluated the performance of machine learning classification models for diabetes prediction. Table I shows the work on the most used classification models in the literature, i.e., DT, NB, k-NN, SVM, and ANN. These models in the literature are evaluated using different datasets and evaluation metrics, making an objective comparison difficult. In addition, to our knowledge, no work in the literature evaluates these models in terms of accuracy, F-measure and execution time. We argue that, for an imbalanced dataset, F-measure should be used. This is because, in an imbalanced dataset, a high accuracy number leads to misinterpretation, as only one class will be detected with high accuracy and the minority class will not be detected. In this paper, we conduct a comparative analysis between the models under study using both evaluation metrics, i.e., accuracy and F-measure. In addition, we compare the execution times of the models’ training and testing.
### TABLE I. EVALUATION OF PAST WORKS ON DT, NB, K-NN, SVM AND ANN MODELS

<table>
<thead>
<tr>
<th>Work</th>
<th>Models evaluated</th>
<th>Dataset used</th>
<th>#features</th>
<th>#observations</th>
<th>Evaluation metrics</th>
</tr>
</thead>
<tbody>
<tr>
<td>[14]</td>
<td>DT, NB, k-NN and SVM</td>
<td>9</td>
<td>768</td>
<td>Accuracy, F-measure, Recall, ROC-AUC and misclassification rate</td>
<td></td>
</tr>
<tr>
<td>[15]</td>
<td>SVM</td>
<td>9</td>
<td>768</td>
<td>Accuracy, precision, recall, F-measure and ROC</td>
<td></td>
</tr>
<tr>
<td>[16]</td>
<td>DT, NB and SVM</td>
<td>9</td>
<td>768</td>
<td>Accuracy, F-measure, precision, recall and ROC</td>
<td></td>
</tr>
<tr>
<td>[17]</td>
<td>DT, k-NN, SVM and ANN</td>
<td>9</td>
<td>768</td>
<td>Accuracy, execution time, recall and error rate</td>
<td></td>
</tr>
<tr>
<td>[18]</td>
<td>DT</td>
<td>9</td>
<td>392</td>
<td>Accuracy, F-measure, precision, recall, kappa statistics and absolute error</td>
<td></td>
</tr>
<tr>
<td>[19]</td>
<td>DT, NB and ANN</td>
<td>9</td>
<td>768</td>
<td>Accuracy and error rate</td>
<td></td>
</tr>
<tr>
<td>[20]</td>
<td>DT, k-NN and SVM</td>
<td>UCI diabetes dataset</td>
<td>9</td>
<td>-</td>
<td>Accuracy, Specificity and sensitivity</td>
</tr>
<tr>
<td>[21]</td>
<td>DT, NB, k-NN and SVM</td>
<td>15</td>
<td>2,500</td>
<td>Accuracy and ROC curve</td>
<td></td>
</tr>
<tr>
<td>[22]</td>
<td>NB</td>
<td>9</td>
<td>-</td>
<td>Accuracy, F-measure, precision and recall</td>
<td></td>
</tr>
<tr>
<td>[23]</td>
<td>SVM</td>
<td>NHANES * [25]</td>
<td>14</td>
<td>-</td>
<td>Sensitivity, specificity, positive and negative predicted values and AROC</td>
</tr>
<tr>
<td>[24]</td>
<td>DT, NB and ANN</td>
<td>Real dataset</td>
<td>12</td>
<td>30,122</td>
<td>Accuracy and ROC curve</td>
</tr>
<tr>
<td>[25]</td>
<td>DT, k-NN, SVM and ANN</td>
<td>13</td>
<td>2,536</td>
<td>Accuracy, sensitivity and specificity</td>
<td></td>
</tr>
<tr>
<td>[26]</td>
<td>DT, k-NN and SVM</td>
<td>9</td>
<td>545</td>
<td>Accuracy, F-measure, sensitivity, specificity, MAE*, RAE*, RRSE* and RMSE*</td>
<td></td>
</tr>
<tr>
<td>[27]</td>
<td>ANN</td>
<td>18</td>
<td>8,640</td>
<td>Specificity, sensitivity, positive and negative predicted values</td>
<td></td>
</tr>
<tr>
<td>[28]</td>
<td>DT and ANN</td>
<td>Real dataset</td>
<td>14</td>
<td>68,994</td>
<td>Accuracy, sensitivity and specificity</td>
</tr>
<tr>
<td>[29]</td>
<td>DT and ANN</td>
<td>PIMA Indian</td>
<td>9</td>
<td>768</td>
<td>Accuracy, sensitivity and specificity</td>
</tr>
<tr>
<td>[30]</td>
<td>DT and ANN</td>
<td>CPCSSN® [32]</td>
<td>9</td>
<td>13,309</td>
<td>Specificity, sensitivity, misclassification rate and Area under the ROC curve (AROC)</td>
</tr>
<tr>
<td>[31]</td>
<td>DT and ANN</td>
<td>CPCSSN® [32]</td>
<td>9</td>
<td>4,678</td>
<td>AROC*</td>
</tr>
</tbody>
</table>

III. MACHINE LEARNING MODELS FOR DIABETES PREDICTION

In this section, we explain the machine learning classification models under study and how we apply them to diabetes prediction.

A. Decision Tree

Decision Tree (DT) uses a tree structure to define the sequences of decisions and their corresponding outcomes [7]. Each feature in the dataset, i.e., a diabetes risk factor is represented by a node as shown in Fig. 1. The top node is called the root node. At each node, the model decides to select a particular branch and traverse down. A node having a class label (diabetic or non-diabetic) is called a leaf node. The model uses a greedy algorithm for the selection of a risk factor to be used for splitting the tree. The most informative risk factor is selected based on the value of information gain.

![Decision Tree Classification Model](image)

**Fig. 1.** Decision Tree Classification Model

B. Naïve Bayes

Naïve Bayes (NB) is based on the Bayes’ theorem that formulates the relationship between the probabilities and conditional probabilities of two events [7]. For an observation with a set of features (risk factors for diabetes), the goal of the model is to predict the class (diabetic or non-diabetic) that maximizes the conditional probability of the class for the given features as shown in Fig. 2. NB assumes that each feature in the dataset is conditionally independent of every other feature for a given class.

![Naive Bayes Model](image)
C. k-Nearest Neighbor

k-Nearest Neighbor (k-NN) does not build a prediction model based on the training dataset. k-NN stores the dataset and classifies a new observation based on how likely that observation is to be a member of either diabetic or non-diabetic class [8]. This is based on the class of the majority of the k nearest observations. The positive integer k is the model parameter that represents the number of nearest neighbors. The nearest neighbors are determined by calculating the distance between the observation whose class has to be predicted and each observation in the training dataset. As shown in Fig. 3 the new observation is assigned to the diabetic class as the majority of the observations among the 5 (value of k) nearest neighbors are diabetic.

D. Support Vector Machine

Support Vector Machine (SVM) creates a decision boundary known as hyperplane that separates the n-dimensional observations into different classes [9]. The hyperplane is created using the extreme points in the dataset that are known as the support vectors. A hyperplane is created in a way that there is a maximum possible margin between the support vectors of the opposite class. In the case of a non-linear dataset such as diabetes, SVM uses a kernel trick to transform the input features space into a higher dimensional space to generate the hyperplane. Fig. 4 shows the SVM hyperplane that separates the diabetic and non-diabetic observations. Fig. 4 is represented only for two risk factors for easy understanding.

E. Artificial Neural Network

Artificial Neural Network (ANN) is an iterative process that consists of a network of neurons [10]. The implementation of ANN in the literature is based on Multilayer Perceptron (MLP). MLP is a feedforward neural network that utilizes backpropagation for training. An MLP consists of three layers: input layer that consists the diabetes risk factors, hidden layer and the output layer that consists of the diabetic and non-diabetic class as shown in Fig. 5. After each iteration, the model calculates the error for each neuron by comparing the result of the output layer and the actual class labels. The error terms are then used to adjust the weights in the hidden layer such that the prediction accuracy increase in the next iteration.

IV. PERFORMANCE ANALYSIS

In this section, we analyze and compare the performance of the most used classification models for diabetes prediction. We evaluate the models in terms of accuracy, F-measure and execution with and without feature selection.
A. Experimental Environment

We evaluate the performance of the classification model using the UCI diabetes dataset [11]. The dataset includes patients’ outcomes for 10 years of clinical care at 130 US hospitals between 1999-2008. It has 55 attributes and 100,000 observations. We evaluate the performance of the classifier with and without feature selection. We use the Correlation Attribute Evaluator [34] feature selection algorithm as it has shown an improvement in the diabetes prediction accuracy for classification models in the literature [21]. The Correlation Attribute Evaluator algorithm evaluates the Pearson’s correlation of each feature and the diabetes class labels and selects the features that have a moderately positive or negative correlation. We use Weka 3.8 [34] for the implementation and evaluation of the studied models.

B. Experiments

1) Data Preprocessing

We first remove the irrelevant features available in the dataset such as encounter id, discharge deposition id, hospital time in and time out, patient number, and payer code. We also remove the feature “weight” as it has 100% missing values. The resultant dataset includes race, gender, age, diagnosis 1, diagnosis 2, diagnosis 3 and diabetes medication. Diagnosis 1, 2 and 3 represent the results of the primary, secondary and additional secondary diagnosis respectively. We then create diabetic and non-diabetic class labels based on the diabetes medication feature available in the dataset. The class value is set to 1, i.e., diabetic, if the corresponding value in the diabetes medication column is ‘yes’, else it is set to ‘0’, i.e., non-diabetic. In addition, we remove all the observations having missing values. For diagnosis 1, 2 and 3, we extract the ICD-9 code values [35] of the diseases that are risk factors of type 2 diabetes such as obesity, hypertension, and cardiovascular disease. A column for each risk factor is then added. The value for every observation for each risk factor is set to ‘1’ if the disease appears in either diagnosis 1, 2 and 3, otherwise, it is set to ‘0’. We convert the numerical range values for the age in categorical values with each category representing an interval of 10 years. Race and gender risk factors are also converted into categories. For instance, a binary category for male and female. The preprocessed dataset contains 12 features and 65,840 observations (51,034 diabetic and 14,806 non-diabetic).

2) Model Building

We evaluate the performance of the studied models with and without feature selection algorithm. The training and testing datasets were obtained using 10-fold cross-validation method. For the k-NN, we run the model for different values of ‘k’ from 1 to square of the number of observations in the dataset and select the value of ‘k’ that gives the highest accuracy. For the SVM model, we implement the polynomial and Radial Basis Function (RBF) kernels. We measure the accuracy, F-measure and execution time for each model. The accuracy and the F-measure are calculated using (1) and (2) respectively.

\[
Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)
\]

\[
F - measure = \frac{2(Recall \times Precision)}{Recall + Precision} \quad (2)
\]

where TP (True Positive) represents the number of observations in the positive class that are predicted as negative, TN (True Negative) represents the number of observations in the negative class predicted as positive, FP (False Positive) represents the number of observations in the negative class predicted as positive and FN (False negative) represents the number of observations in the positive class predicted as negative. The recall and precision for the positive (negative) class are calculated using (3) and (4) respectively.

\[
recall = \frac{TP(TN)}{TP(TN) + FN(FP)} \quad (3)
\]

\[
Precision = \frac{TP(TN)}{TP(TN) + FP(FP)} \quad (4)
\]

C. Experimental Results Analysis

Fig. 6 shows the accuracy and F-measure of the models under study for the UCI diabetes dataset without feature selection. The accuracy of all the models is almost the same. However, the DT and SVM models have no F-measure values even though they have almost the same accuracy as the other models under study. This is because these models classify all the observations in the testing dataset as diabetics and are not able to detect the non-diabetic class. The reason for this is that the dataset used in the experiments is unbalanced, i.e., 77.5% of the observations are diabetic and 22.5% of them are non-diabetic. Consequently, our experimental results reveal that it is important to evaluate the models in terms of F-measure as accuracy alone can be misleading. Comparing the execution times of the models without feature, the execution time of NB is the least among the models having an F-measure value (Table II). Consequently, based on our experimental results, NB is the most efficient model in terms of accuracy, F-measure and execution time.

<table>
<thead>
<tr>
<th>Model</th>
<th>Execution time (minutes)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Without feature selection</td>
</tr>
<tr>
<td>DT</td>
<td>0.2166</td>
</tr>
<tr>
<td>NB</td>
<td>0.0166</td>
</tr>
<tr>
<td>k-NN</td>
<td>4.65</td>
</tr>
<tr>
<td>SVM (polynomial)</td>
<td>37.2166</td>
</tr>
<tr>
<td>SVM (RBF)</td>
<td>574.6166</td>
</tr>
<tr>
<td>ANN</td>
<td>12.9166</td>
</tr>
</tbody>
</table>

Fig. 6. Accuracy and F-measure of the Models without Feature Selection
The feature selection algorithm in our experiments extracts 6 risk factors for the prediction of diabetes type 2. These risk factors are gender, age, blood pressure, cholesterol, heart disease, and obesity. Fig. 7 shows the performance of the studied models in terms of accuracy and F-measure with feature selection. It shows that the accuracy of the models with feature selection (Fig. 7) is nearly the same as that without feature selection (Fig. 6). This reveals that the risk factors selected by the feature selection algorithm are the most significant for the prediction of diabetes type 2. This is also confirmed by the statistical and clinical results in the literature [36, 37]. However, the DT, SVM and k-NN models have no F-measure value as they are not able to detect the minority non-diabetic class.

The execution time of the models with feature selection is shown in Table II. The time with feature selection is less compared to that without feature selection because of the reduced number of features in the dataset. However, for k-NN, this reduction of time is not beneficial as the model is not able to detect the minority class after removing features. For the models having an F-measure value with feature selection, NB has the least execution time (Table II).

In summary, NB outperforms all the other studied models for the prediction of diabetes type 2. NB gives an accuracy of 77.5% with and without feature selection. The execution time of NB with and without feature selection is the same, i.e., 0.016 minutes.
V. CONCLUSIONS

Diabetes is a global crisis that is primarily driven by rapid urbanization, changing lifestyles, and uneven dietary patterns. It is crucial to predict the prevalence of diabetes in an individual to reduce the risk of diabetes development and save lives. Machine learning classification models is a promising intelligent approach to predict medical diseases, such as diabetes.

In this paper, we evaluate and compare the performance of the most used machine learning classification models: DT, NB, k-NN, SVM and ANN. We conduct a comparative analysis of the models in terms of accuracy, F-measure and execution time with and without feature selection. This is using the UCI diabetes dataset which has an important number of observations. Our experimental results show that NB is the most accurate model to predict diabetes mellitus type 2 and has the least execution time, with and without feature selection. Our results reveal that evaluating the performance of the classification models using only accuracy measure can be misleading. This is because, in the case of an imbalanced dataset, which is frequent in the health domain, the model might have high accuracy but will not be able to predict the minority class. Consequently, it
is important to include F-measure as one of the evaluation metrics. Our experiments show that feature selection reduces the model execution time and gives insights on the most significant features that are necessary for prediction without accuracy degradation. They reveal the most accurate models that can detect both diabetic and non-diabetic classes even with feature selection. Based on our experimental results, the most significant risk factors for the prediction of diabetes type 2 are gender, age, blood pressure, cholesterol, heart disease and obesity.

REFERENCES