Enhancing Diabetes Detection using Machine Learning Models

Prof. RAJAKUMAR. B1, VISHAL. P2, TANUJA. V3, AQUVINA. A4, BHARATH.K5

1,2,3,4,5 DEPARTMENT OF ARTIFICIAL INTELLIGENCE & DATA SCIENCE
J.N.N INSTITUTE OF ENGINEERING
Kannigapair, Thiruvallur, India

Abstract: This research aims to develop and evaluate machine learning models to predict the onset of diabetes using the Pima Indians Diabetes Dataset. Diabetes is a prevalent chronic condition posing significant health risks if not detected early. Diabetes is a disease caused due to the increase level of blood glucose. This high blood glucose produces the symptoms of frequent urination, increased thirst, and increased hunger. Diabetes is one of the leading cause of blindness, kidney failure, amputations, heart failure and stroke. When we eat, our body turns food into sugars, or glucose. At that point, our pancreas is supposed to release insulin. Insulin serves as a key to open our cells, to allow the glucose to enter and allow us to use the glucose for energy. But with diabetes, this system does not work. Type 1 and type 2 diabetes are the most common forms of the disease, but there are also other kinds, such as gestational diabetes, which occurs during pregnancy, as well as other forms. Machine learning is an emerging scientific field in data science dealing with the ways in which machines learn from experience. By comparing various algorithms, including Logistic Regression, Decision Trees, Random Forest, Support Vector Machines (SVM), and K-Nearest Neighbors (KNN), we aim to identify the most effective model for diabetes prediction. Our evaluation metrics include accuracy, precision, recall, F1-Score, and ROC-AUC.

I. INTRODUCTION
Diabetes is a chronic metabolic disorder characterized by elevated blood sugar levels, which, if left untreated, can lead to severe complications such as cardiovascular diseases, nerve damage, kidney failure, and vision impairment. According to the World Health Organization (WHO), diabetes affects approximately 422 million people worldwide, with a rising prevalence that poses a significant public health challenge. Early detection and effective management of diabetes are crucial in mitigating its impact on individuals and healthcare systems.

Machine learning, a subset of artificial intelligence, offers promising solutions for the early prediction and diagnosis of diabetes. By leveraging large datasets and advanced algorithms, machine learning models can identify patterns and correlations within medical data that are often imperceptible to human analysis. These predictive models can provide healthcare professionals with valuable insights for timely intervention, personalized treatment plans, and improved patient outcomes.

Types of Diabetes:

Type 1 diabetes means that the immune system is compromised and the cells fail to produce insulin in sufficient amounts. There are no eloquent studies that prove the causes of type 1 diabetes and there are currently no known methods of prevention.

Type 2 diabetes means that the cells produce a low quantity of insulin or the body can’t use the insulin correctly. This is the most common type of diabetes, thus affecting 90% of persons diagnosed with diabetes. It is caused by both genetic factors and the manner of living.

Gestational diabetes appears in pregnant women who suddenly develop high blood sugar. In two thirds of the cases, it will reappear during subsequent pregnancies. There is a great chance that type 1 or type 2 diabetes will occur after a pregnancy affected by gestational diabetes.

Causes of diabetes:

One of the primary factors is genetics; a family history of diabetes significantly increases an individual’s risk. Lifestyle choices, such as poor diet and physical inactivity, also play a crucial role. A diet high in refined sugars and
unhealthy fats can lead to obesity, a major risk factor for Type 2 diabetes. Additionally, a sedentary lifestyle contributes to weight gain and insulin resistance. Other factors include age, with the risk increasing as people get older, and certain ethnic backgrounds, where some populations are more predisposed to developing diabetes. Moreover, conditions such as hypertension and high cholesterol are often associated with diabetes, forming part of the metabolic syndrome. Environmental factors and stress can exacerbate these risks, while gestational diabetes during pregnancy increases the likelihood of developing Type 2 diabetes later in life. Understanding these causes is essential for prevention and management strategies.

II. LITERATURE SURVEY
Numerous studies have explored the application of machine learning in diabetes prediction.

1. Mitush Soni proposed machine learning algorithms for providing better accuracy in diabetes prediction. In this paper, the dataset contains 500 negative outcomes means no diabetes and 268 positive outcomes means. Diabetes and for predicting accurately they have used 6 machine learning algorithms and among these 6 algorithms random forest algorithm predicts with 77% accuracy.

2. Kumari and Chitra used SVM, RFC, DT, MLP, and LR, as well as four k-fold cross-validations (k = 2, 4, 5, 10) in their research. According to the researchers, MLP with four-fold cross-validation achieves the best accuracy, at 78.7%. They discovered that MLP outscored all other algorithms.


4. Shafi et al. [13] reported that because diabetes is a serious illness, early detection is always a struggle. This study used machine learning classification methods to develop a model that could solve any problem and that could be used to identify diabetes development early on. The authors of this research made concerted efforts to develop a framework that could accurately predict the likelihood of diabetes in patients. As part of this study, the three ML approach classification algorithms—DT, SVM, and NBC—were studied and assessed on various measures. In the study, the PID data set acquired from the UCI repository was used to save time and produce precise findings. The experimental results suggested that the NBC approach was adequate, with a 74% accuracy, followed by SVM with a 63% accuracy and the DT with a 72% accuracy.

III. CORPUS
Creating a corpus for diabetes prediction using machine learning involves compiling relevant datasets that contain features related to diabetes and its predictors. Here’s a curated list of datasets that you can use:

1. **Pima Indians Diabetes Database:**
   - **Description:** Contains data on female Pima Indian population near Phoenix, Arizona, USA. This dataset is commonly used for diabetes prediction studies.
   - **Features:** Age, number of pregnancies, glucose level, blood pressure, skin thickness, insulin level, BMI, diabetes pedigree function, and diabetes outcome (1 for diabetic, 0 for non-diabetic)

2. **Indian Diabetes Dataset (IDDS):**
   - **Description:** Dataset containing health parameters of individuals from India.
   - **Features:** Age, gender, BMI, blood pressure, and various other health indicators.

3. **UCI Diabetes Dataset:**
Description: Available from the UCI Machine Learning Repository, this dataset provides information on diabetes patients.
Features: Age, sex, body mass index (BMI), average blood pressure, and six blood serum measurements.

4. **Diabetes 130-US hospitals for years 1999-2008 Data Set:**
   - Description: Contains over 100,000 diabetes patient records from 130 US hospitals.
   - Features: Demographic information, diagnoses, medications, and medical test results.

**IV. PROPOSED MODEL**

Goal of the paper is to investigate for model to predict diabetes with better accuracy. We experimented with different classification and ensemble algorithms to predict diabetes. In the following, we briefly discuss the phase.

- **Dataset Description:**
  - The data is gathered from UCI repository which is named as Pima Indian Diabetes Dataset. The dataset have many attributes of 768 patients.
  - The diabetes data set consists of 2000 data points, with 9 features each.
  - “Outcome” is the feature we are going to predict, 0 means No diabetes, 1 means diabetes.
  - visualizes the correlation of the features using a seaborn heatmap.

- **Data Preprocessing:**
  - Data preprocessing is most important process. Mostly healthcare related data contains missing vale and other impurities that can cause effectiveness of data. To improve quality and effectiveness obtained after mining process, Data preprocessing is done. To use Machine Learning Techniques on the data set effectively this process is essential for accurate result and successful prediction. The above fig shows the Shows that all the columns are non-null.

- **Feature Selection**
  - The number of features in the dataset is not large. So, in the first round of this study, we performed our machine learning experiment using all eight predictors: pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, diabetes pedigree function, and age.
  - In the second round of this study, we used SelectKBest feature selection method, provided by Scikit-learn, to extract the best features in the dataset. The SelectKBest removes all but k highest scores features (Scikit-learn.feature_selection.SelectKBest Scikit-Learn 1.1.3 Documentation, 2011). The classification scoring function used is the “f classif” function, which returns the ANOVA F-value between the label/feature for classification tasks (Scikit-learn.feature_selection.f_classif Scikit-Learn 1.1.3 Documentation, 2011). The F-value and p-value of each of the predictors were used to identify the best risk factors for the prediction of diabetes mellitus.

- **Apply Machine Learning**
  - When data has been ready we apply Machine Learning Technique. We use different classification and ensemble techniques, to predict diabetes. The methods applied on Pima Indians diabetes dataset. Main objective to
apply Machine Learning Techniques to analyze the performance of these methods and find accuracy of them, and also been able to figure out the responsible/important feature which play a major role in prediction. The Techniques are follows

- **k-Nearest Neighbors:**
  The k-NN algorithm is arguably the simplest machine learning algorithm. Building the model consists only of storing the training data set. To make a prediction for a new data point, the algorithm finds the closest data points in the training data set, its "nearest neighbors."

  First, let’s investigate whether we can confirm the connection between model complexity and accuracy:

  ![Training and Testing Accuracy Graph](image)

  ![Table 1](image)

<table>
<thead>
<tr>
<th>Training Accuracy</th>
<th>Testing Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.81</td>
<td>0.78</td>
</tr>
</tbody>
</table>

- **Logistic Regression:**
  Logistic Regression is one of the most common classification algorithms.
  
  → In first row, the default value of C-1 provides with 77% accuracy on the training and 78% accuracy on the test set.
  
  → In second row, using C-0.01 results are 78% accuracy on both the training and the test sets.

  ![Coefficient Magnitude](image)

  → Using C-100 results in a little hit lower accuracy on the training set and little hit highest accuracy on the test set, confirming that less regularization and a more complex model may not generalize better than default setting.

- **Support Vector Machine:**
  This classifier aims at forming a hyper plane that can separate the classes as much as possible by adjusting the distance between the data points and the hyper plane. There are several kernels based on which the hyper plane is decided. I tried four kernels namely, linear, poly, rbf, and sigmoid.
Accuracy Comparison:

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Training Accuracy</th>
<th>Testing Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>k-Nearest Neighbors</td>
<td>85%</td>
<td>50%</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>70%</td>
<td>50%</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>85%</td>
<td>85%</td>
</tr>
<tr>
<td>Random Forest</td>
<td>70%</td>
<td>50%</td>
</tr>
<tr>
<td>SVM</td>
<td>70%</td>
<td>70%</td>
</tr>
</tbody>
</table>

V. IMPLEMENTATION

Data Loading and Exploration:

Import pandas as pd
#read the file
Df = pd.read_csv("diabetes prediction.csv")
Df.head()
Df.shape
Df.describe()

Data Preprocessing:

X=diabetes_dataset.drop(columns ‘Outcome’, axis 1)
Y=diabetes_dataset[‘Outcome’]
Print(X)
Print(Y)

Model Training:

X_train,X_test, Y_train, Y_test train_test_split(X, Y, test_size=0.2, stratify=Y, random_state=2) print(X.shape,
X_train.shape,X_test.shape)
X_train_prediction classifier.predict(X_train)
training_data_accuracy= accuracy_score(X_train_prediction, Y_train) print('Accuracy score of the training data:',
training_data_accuracy)
X_test_prediction classifier.predict(X_test)
test_data_accuracy accuracy_score(X_test_prediction, Y_test)
print('Accuracy score of the test data:', test_data_accuracy)

filename ‘trained_model.sav
pickle.dump(classifier, open(filename, ‘wb’))

Model Evaluate:

A=int(input("pregnancies: “))
B=int(input("Glucose: “))
D= int(input("SkinThickness: “))
E=int(input("Insulin: “))
C = int(input("BloodPressure: “))
F = float(input("BMI: “))
G = float(input("Diabetes PedigreeFunction: “))
input_data_str = [a,b,c,d,e,f,g,h]
input_data_as_numpy_array = np.asarray(input_data_str)
input_data_reshaped = array.reshape(1,-1)
prediction.classifier.predict(input_data_reshaped)
print(prediction)
if(prediction[0] == 0):
    print("The person is not diabetic")
else:
    print("the person is diabetic")

Outcome:

Bar diagram

V. CONCLUSION

In this research, we have explored the application of machine learning algorithms for predicting diabetes, a critical health condition that requires early detection and effective management. By employing various classification algorithms, including Decision Trees (DT), Support Vector Machines (SVM), and Naïve Bayes Classifier (NBC), we aimed to develop a predictive model capable of accurately identifying individuals at risk of diabetes.

The research highlights the significant role that machine learning can play in early diabetes detection. By leveraging data-driven approaches, healthcare providers can potentially identify at-risk individuals sooner, allowing for timely interventions that can prevent or delay the onset of diabetes-related complications.

Future work can expand upon this study by incorporating larger and more diverse datasets, exploring additional machine learning techniques, and integrating feature engineering to enhance model performance. Furthermore, the developed framework and classifiers can be adapted to predict other diseases, contributing to broader advancements in medical diagnostics and personalized healthcare.

Overall, the study reinforces the promise of machine learning in transforming diabetes prediction and management, ultimately leading to better health outcomes for patients.

VI. REFERENCE


