Report

DIABETIC RISK ASSESSMENT USING A NEURAL NETWORK

Ehtisham Ur Rehman
ehtishamrehman@uetpeshawar.edu.pk

About the dataset

The dataset used in this study is the PIMA Indians diabetes database, which is an open-source dataset. This dataset was initially collected by the National Institute of Diabetes and Digestive and Kidney Diseases. The purpose of this dataset is to predict whether or not a patient has diabetes based on certain diagnostic measurements included in the dataset. The dataset includes multiple medical predictor variables and a single target variable. Variables used as predictors include the patient's age, BMI, insulin level, blood pressure, skin thickness and diabetes pedigree function and the number of pregnancies.

Contents

1. Introduction

The term "diabetes" refers to a group of metabolic disorders that are characterized by elevated levels of glucose (BG) in the blood [1]. Long-term complications of diabetes can include heart disease, kidney failure, circulatory disorders, and nerve damage [2, 3]. According to recent survey, there will be approximately 642 million people living with diabetes all over the world by the year 2040 [4]. Diabetes can generally be broken down into these three sub-types: type I diabetes (also known as T1D), type II diabetes (also known as T2D), and gestational diabetes [5]. Young adults under the age of 30 are the ones who are diagnosed with type 1 diabetes the most frequently. T1D can cause several symptoms, including polyuria, thirst, constant hunger, weight loss, changes in vision, and fatigue [6]. T2D affects adults over the age of 45 on average and is frequently associated with obesity, hypertension, dyslipidemia, arteriosclerosis, and a variety of other diseases [7]. The third form of diabetes that can occur in pregnant women is referred to as gestational diabetes.

Blood glucose concentrations should be kept throughout the day within a healthy range of 70-180 mg/dL. This is true even though BG levels are subject to significant fluctuations due to things like meals, exercise, stress, and other factors. In order to keep blood glucose levels and insulin production at normal levels, the body needs to be able to maintain normal glucose levels. Both hypoglycemic and hyperglycemic episodes are potential outcomes of BG and insulin levels that are out of whack with one another. Blood sugar levels that are below 70 mg/dL are considered to be hypoglycemic, while blood sugar levels that are above 180 mg/dL are considered to be hyperglycemic. Both hypoglycemia and hyperglycemia are further subdivided into two categories, referred to as levels 1 and 2. If the blood glucose levels fall below 70 mg/dL, the person is said to be experiencing level 1 hypoglycemia; if they fall below 54 mg/dL, the person is said to be experiencing level 2 hypoglycemia [8]. In severe cases, hypoglycemia can result in unconsciousness, confusion, muscle spasms, and even death [9]. This condition is one of the most dangerous complications that can arise from diabetes. In a similar manner, level 1 hyperglycemia is defined as having BG levels that are greater than 180 to 250 mg/dL, and level 2 hyperglycemia is defined as having BG levels that are greater than 250 mg/dL [10]. Long-term complications of untreated hyperglycemia include cardiovascular disease, nerve damage (neuropathy), ketoacidosis, diabetic retinopathy, damaged nerves, or poor blood flow, any of which can result in severe skin infections, ulcerations, and, in extreme cases, amputation [11].

Deep learning is a sub-field of artificial intelligence that attempts to solve extremely difficult problems by modeling their solutions after how the human brain processes information. During the process of deep learning, we make use of neural networks, which are comprised of multiple operators that are positioned within nodes [12]. These operators work together to assist in segmenting the problem into more manageable pieces that are then individually solved. However, the implementation of neural networks can be quite challenging. Keras, a framework for deep learning, is able to solve this issue without any difficulty.

Keras is a high-level application programming interface (API) for deep learning that was developed by Google. It is used to implement neural networks [13]. It is a neural network implementation tool that is written in Python and is intended to simplify the process of deploying neural networks. Additionally, it supports the computation of multiple back-end neural networks. Because it offers a Python front-end with a high level of abstraction and provides a choice between multiple back-ends for computation, learning Keras and working with it is not too difficult.

Using a neural network system, we searched for indicators of future diabetes cases in Pima Indian women [14]. The Pima Indian Women's Diabetes Database, maintained by the National Institute of Diabetes and Digestive and Kidney Diseases and funded by the National Institutes of Health, was utilized for this study. The dataset is accessible to anyone interested on the Kaggle platform. The neural network constructed with Keras is used to assess diabetes risk.

2. Methodology

2.1 Dataset

The National Institute of Diabetes, Digestive, and Kidney Diseases supported and published the dataset referred to as the Pima Indian Diabetes database. The dataset is accessible to the public on the Kaggle website; it is an open-source dataset containing patient records for female patients. Each case represents a female Pima Indian participant. The dataset contains 768 cases. Each case has a binary indicator, non-diabetic denoted by 0 or diabetic denoted by 1. The data set consists of 500 non-diabetic cases and 268 diabetic cases. In addition, the dataset includes eight distinct variables or features that can serve as predictors of our binary dependent variable (diabetes or non-diabetes). The dataset consists of the following features:

- **Pregnancies:** This variable represents the number of pregnancies a female Pima Indian experienced.
- **Glucose Level:** It measures the plasma glucose concentration over 2 hours in an oral glucose tolerance test.
- **Blood Pressure:** Each patient diastolic blood pressure is recorded, while the metric used in the dataset is (mm Hg).
- **Skin Thickness:** The metric used in the dataset for this variable is triceps skinfold thickness (mm).
- **Insulin:** Insulin in this dataset corresponds to 2-hour serum insulin. The unit used by Pima Indians to measure the two-hour serum insulin concentration (mu U/ml).
- **Body mass index (BMI):** It is a measure of obesity and health, and is frequently employed in statistical analysis.
- **Diabetes Pedigree Function:** Diabetes risk due to family history is reflected by the DBF variable.
- **Age:** The range of ages in the dataset is between 21 and 81 years.

- **Outcome:** Classification variable in which 0 indicates a female does not have diabetes and 1 indicates the participant has diabetes.

The range and statistical summary of the dataset before preprocessing is mentioned in table 1.

Table 1: Summary of the dataset before preprocessing

	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age
Count	768	768	768	768	768	768	768	768
Mean	3.85	120.89	69.11	20.54	79.80	31.99	0.47	33.24
STD	3.37	31.97	19.36	15.95	115.24	7.88	0.33	11.76
Min value	0	0	0	0	0	0	0.08	21.00
IQR 25%	1.00	99.00	62.00	0	0	27.30	0.24	24.00
IQR 50%	3.00	117.00	72.00	23.00	30.50	32.00	0.37	29.00
IQR 75%	6.00	140.25	80.00	32.00	127.25	36.60	0.63	41.00
Max value	17.00	199.00	122.00	99.00	846.00	67.10	2.42	81.00

2.2 Data preprocessing

The dataset has missing values, as shown in the table 1, which may affect the model's training. Thus, data preprocessing was performed to guarantee that a clean dataset was provided to the neural network for optimal performance. For features like ”Glucose,” ”Blood Pressure,” ”Skin Thickness,” ”Insulin,” and ”Body Mass Index,” missing data treatment was performed. After separating the data into a training (80%) and testing set (20%), zeros were replaced using a simple average to remove any potential bias. Each feature's average was determined independently and re-entered into the same column.

2.3 Neural Network

As shown in figure 1, the neural network consists of three layers, and each layer contains a different number of neurons: 12, 8, and 1, respectively. Additionally, the checkpoint and saving weights of the model were used to save the model after certain epochs had been completed.

The eight features from the dataset were used as an input while the outcome was used as an output to the neural network. the total number of epochs used were 1000 while the batch size was 16. The model was saved once the validation accuracy of the model was improved in order to have a checkpoint.

3. Results

It is just as important to evaluate a machine learning model as it is to build one. As a result, the following section discusses the various metrics that are used to evaluate our model performance.

3.1 Performance metrics

When evaluating the models, many different metrics are used. Nevertheless, a confusion matrix is not a metric that is used to evaluate a model; rather, it is a tool that

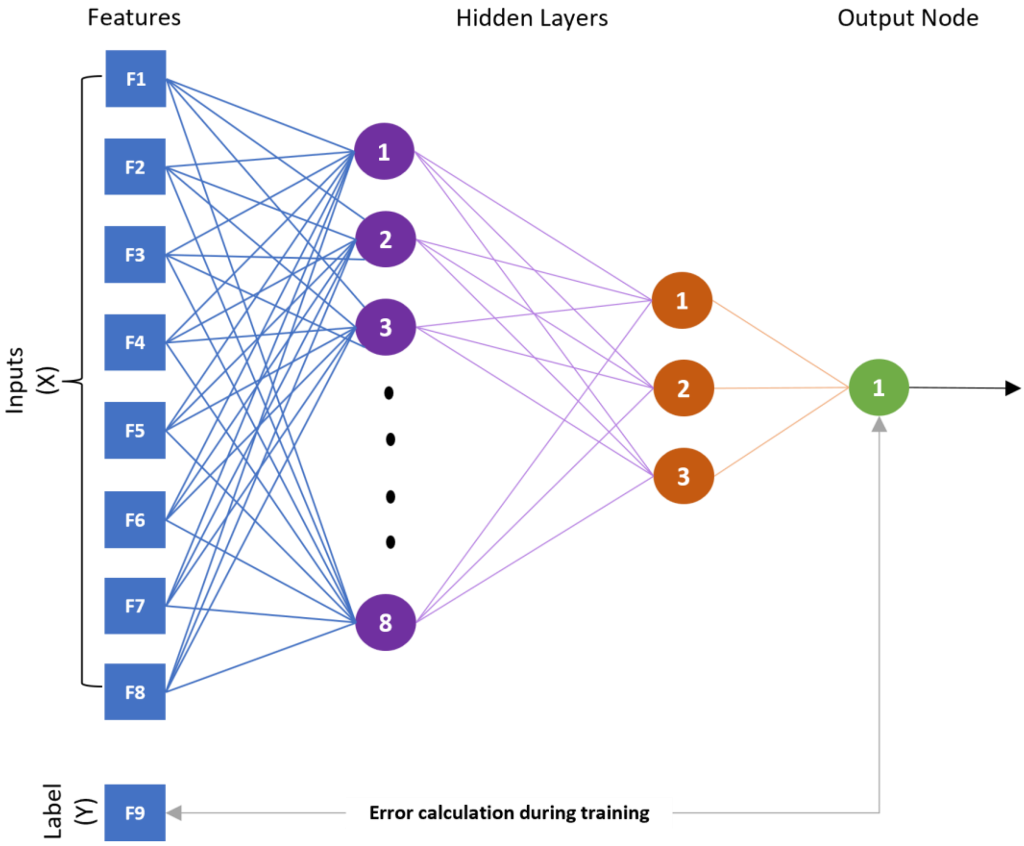

Figure 1: 3-layered neural network

offers insight into the predictions. It is necessary to perform the calculation of the confusion matrix in order to have an understanding of other classification metrics such as accuracy, precision, and recall. The confusion matrix takes things a step further than simple classification accuracy by displaying, for each class, both correct and incorrect (that is, true and false) predictions. Figure 2 shows the confusion matrix of the model.

- **TP (True Positive):** The model predicted that it would be positive, and it was. In our case, the model said that the patient would have diabetes, and they do.
- **TN (True negative):** The model said negative, which is what happened, which means that in our project, the patient does not have diabetes, as the model predicted, and they do not have diabetes.
- **FP (False Positive):** The model predicted that the result would be positive, but it was wrong. In our case, the model said that the patient has diabetes, but they are not.
- **FN (False Negative):** The model predicted that the answer would be negative, which is false. In our case, the model said a patient would not have diabetes, but they do.

TP, TN, FP, and FN are used to figure out the classification method's performance

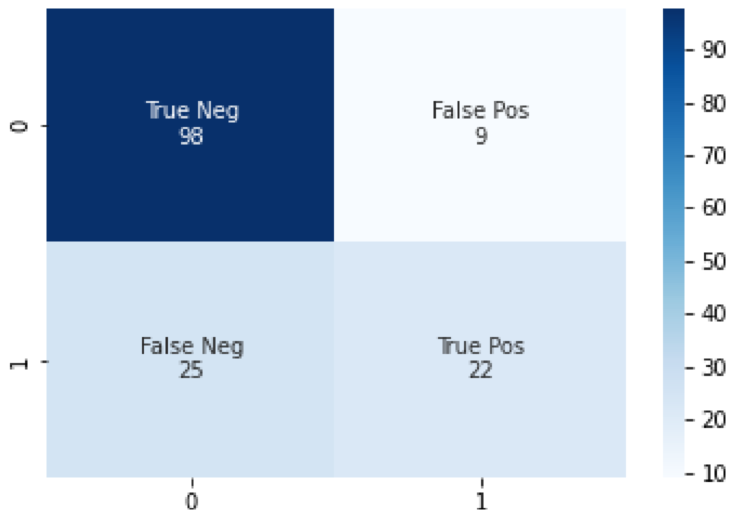

Figure 2: Confusion matrix

measurement or metric. To choose the right machine learning model and make decisions based on the predictions, it is important to understand the different relevance metrics. The following measures were used to evaluate how effectively the ML model worked;

- **Accuracy:** It is the number of accurate predictions divided by the total number of predictions. Our model accuracy is 77.92%. Mathematically,

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FN + FP)}$$

- **Precision:** A model's precision is the proportion of true positives to predicted positive observations. The percentage of patients we correctly identified as diabetic in our study. Our model precision is 70.96%. Mathematically,

$$Precision = \frac{TP}{(TP + FP)}$$

- **Mathew Cross-Correlation (MCC):** MCC determines the correlation between the true classes and the predicted labels, was used to evaluate the performance of the implemented model. MCC accepts values between -1 and 1. A classification value of -1 shows that the model predicts the positive and negative of the target variable oppositely. If MCC is set to 0, the model predicts positive and negative classes at random. The model is ideal for value 1. The MCC of our system is 0.44. Mathematically,

$$MCC = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

4. Discussion

Type 1 diabetes, also known as juvenile diabetes, is a chronic autoimmune condition that causes the body to be unable to produce insulin, the hormone that regulates blood sugar. Our bodies cannot use the sugar in our bloodstream as energy without insulin, resulting in diabetic ketoacidosis. This project aimed to identify factors and indicators that have a high risk of predicting Type I Diabetes in women, allowing those women to take the necessary precautions to avoid the disease's onset.

In this work, we predicted diabetes using the female Pima Indians diabetes dataset. We used glucose, pregnancies, age, insulin, and BMI to predict the outcome. A neural network having three layers, was used to predict diabetes in a patient. The algorithm was trained with 80% data as training while 20% was reserved for testing. From figures 3 and 4, it is evident that after 500 epochs the system accuracy and loss didn't change.

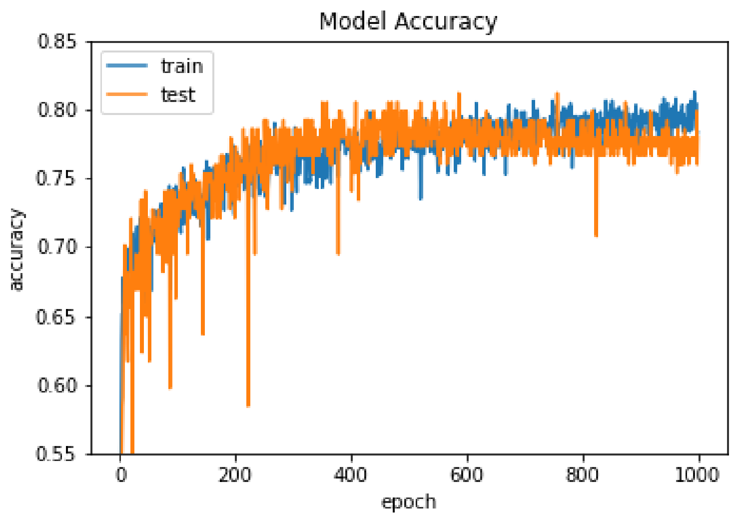

Figure 3: Model accuracy during each epoch

The model was assessed using metrics such as accuracy, precision, and MCC. The analysis results are promising, but due to data constraints, more than 768 samples in the dataset are required for the ML model. If this project were to be expanded in the future, it would be prudent to collect more data, as the dataset only contained 798 records. We could also include more features, such as daily exercise and dietary records. Having more and different data may help better predict diabetes, allowing both men and women to take the necessary steps to prevent the disease's onset.

5. Conclusion

A binary classification model is implemented using a neural network model. PIMA Indian diabetes dataset is used to train and test the model performance. The results

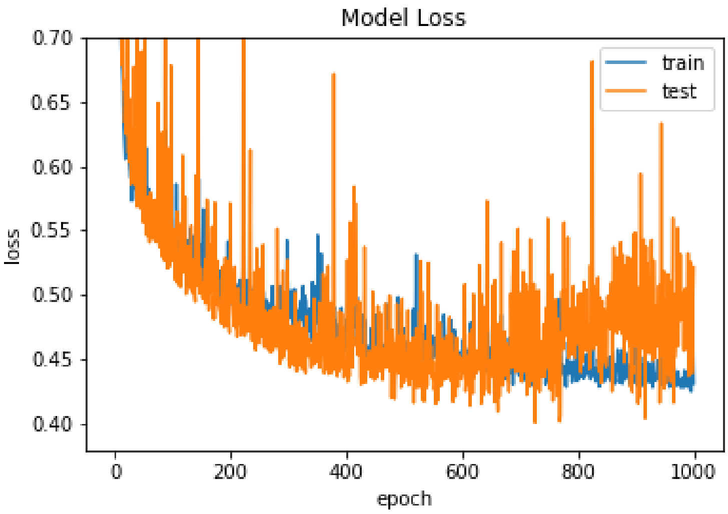

Figure 4: Model loss during each epoch

shows that the model can efficiently predict the diabetes, however, for more efficiently large amount of data is required for training the model.

Bibliography

[1] A. D. Association, “Diagnosis and classification of diabetes mellitus,” *Diabetes care*, vol. 37, no. Supplement 1, pp. S81–S90, 2014.

[2] A. Krasteva, V. Panov, A. Krasteva, A. Kisselova, and Z. Krastev, “Oral cavity and systemic diseases—diabetes mellitus,” *Biotechnology & Biotechnological Equipment*, vol. 25, no. 1, pp. 2183–2186, 2011.

[3] D. M. Nathan, “Long-term complications of diabetes mellitus,” *New England journal of medicine*, vol. 328, no. 23, pp. 1676–1685, 1993.

[4] P. Zimmet, K. G. Alberti, D. J. Magliano, and P. H. Bennett, “Diabetes mellitus statistics on prevalence and mortality: facts and fallacies,” *Nature Reviews Endocrinology*, vol. 12, no. 10, pp. 616–622, 2016.

[5] G. Danaei, M. Finucane, Y. Lu, G. Singh, M. Cowan, C. Paciorek, J. Lin, F. Farzadfar, Y.-H. Khang, G. Stevens, M. Rao, M. Ali, L. Riley, C. Robinson, and M. Ezzati, “National, regional, and global trends in fasting plasma glucose and diabetes prevalence since 1980: Systematic analysis of health examination surveys and epidemiological studies with 370 country-years and 2 - 7 million participants,” *Lancet*, vol. 378, pp. 31–40, 06 2011.

[6] I. Iancu, M. Mota, and E. Iancu, “Method for the analysing of blood glucose dynamics in diabetes mellitus patients,” in *2008 IEEE international conference on automation, quality and testing, robotics*, vol. 3, pp. 60–65, IEEE, 2008.

[7] G. Robertson, E. D. Lehmann, W. Sandham, and D. Hamilton, “Blood glucose prediction using artificial neural networks trained with the AIDA diabetes simulator: a proof-of-concept pilot study,” *Journal of Electrical and Computer Engineering*, vol. 2011, 2011.

[8] J. Vehi, O. Mujahid, and I. Contreras, “Artificial intelligence and machine learning for diabetes decision support,” in *Advanced Bioscience and Biosystems for Detection and Management of Diabetes*, pp. 259–272, Springer, 2022.

[9] M. Alsahli and J. E. Gerich, “Hypoglycemia,” *Endocrinology and Metabolism Clinics*, vol. 42, no. 4, pp. 657–676, 2013.

[10] R. I. Holt, J. H. DeVries, A. Hess-Fischl, I. B. Hirsch, M. S. Kirkman, T. Klupa, B. Ludwig, K. Nørgaard, J. Pettus, E. Renard, *et al.*, “The management of type 1 diabetes in adults. a consensus report by the american diabetes association (ADA) and the european association for the study of diabetes (EASD),” *Diabetes Care*, vol. 44, no. 11, pp. 2589–2625, 2021.

[11] E. H. Yayan, M. Zengin, A. Akıncı, *et al.*, “The relationship between the quality of life and depression levels of young people with type i diabetes.,” *Perspectives in Psychiatric Care*, vol. 55, no. 2, pp. 291–299, 2019.

[12] J. Noguer, I. Contreras, O. Mujahid, A. Beneyto, and J. Vehi, “Generation of individualized synthetic data for augmentation of the type 1 diabetes data sets using deep learning models,” *Sensors*, vol. 22, no. 13, p. 4944, 2022.

[13] N. Ketkar and N. Ketkar, “Introduction to keras,” *Deep learning with python: a hands-on introduction*, pp. 97–111, 2017.

[14] V. Chang, J. Bailey, Q. A. Xu, and Z. Sun, “Pima indians diabetes mellitus classification based on machine learning (ml) algorithms,” *Neural Computing and Applications*, pp. 1–17, 2022.