Perceptron-Based Prediction of Diabetes Using the Pima Indian Dataset

First Author
Institution1
Institution1 address
firstauthor@i1.org

Abstract

The diagnosis of diabetes is a significant medical challenge that has been the focus of various machine learning techniques. In this work, we implement a Perceptron-based neural network for the prediction of diabetes using the Pima Indian dataset. The dataset consists of several physiological attributes, including glucose levels, body mass index (BMI), and insulin, which are used to predict the binary outcome (diabetic or non-diabetic). We preprocess the data by handling missing values, feature scaling, and creating engineered features. We then train a Perceptron model and evaluate its performance using metrics such as accuracy and the area under the ROC curve (AUC). The experimental results show that our method achieves promising results, demonstrating the efficacy of the Perceptron algorithm in medical diagnosis tasks. Future improvements could include optimizing the model architecture and exploring advanced deep learning approaches. The implementation of this homework can be found on `https://github.com/yyaaoo33/deeplearning-ass1.git`.

1. Introduction

Diabetes is a chronic medical condition affecting millions of people worldwide, with severe long-term complications if not properly managed. Accurate and early diagnosis of diabetes is crucial for effective treatment and prevention of complications. Machine learning techniques have increasingly been applied to predict diabetes using clinical data [8]. The Pima Indian Diabetes dataset, originally from the National Institute of Diabetes and Digestive and Kidney Diseases, provides a well-known benchmark for developing such predictive models.

The Perceptron algorithm, one of the earliest neural network models, has been widely used for binary classification tasks [7]. The Perceptron works by adjusting its weights iteratively based on the classification error, making it a suitable candidate for predicting whether a patient has diabetes. In this paper, we propose a multi-layer Perceptron model to predict the onset of diabetes in patients based on physiological data. We preprocess the data, train the Perceptron model, and evaluate its performance using various metrics, including accuracy and the receiver operating characteristic (ROC) curve.

Our goal is to demonstrate the effectiveness of a Perceptron-based neural network in classifying diabetes outcomes, and to provide a comprehensive analysis of the model's strengths and weaknesses.

2. Related Works

Several approaches have been applied to diabetes prediction, leveraging both traditional statistical methods and modern machine learning techniques. Early works used logistic regression to model the relationship between patient features and diabetes outcomes [6]. However, the rise of machine learning algorithms, such as decision trees, support vector machines (SVMs), and neural networks, has shown promise in improving prediction accuracy [1].

The Perceptron algorithm, introduced by Rosenblatt in 1958, is a foundational element of neural networks and has been widely studied for its capability in binary classification tasks [7]. Modern extensions, such as multi-layer Perceptrons (MLPs), are particularly useful for handling more complex, non-linear relationships between input features and output labels, as demonstrated by recent applications in medical diagnosis [5].

In recent years, neural networks have been effectively applied to the Pima Indian Diabetes dataset. Works by [3] and [4] have demonstrated the usefulness of artificial neural networks and deep learning models in predicting diabetes, often outperforming classical machine learning methods. Our work builds on these efforts by implementing a simpler, yet effective, multi-layer Perceptron to predict diabetes with the goal of understanding its practical applicability and performance.

3. Methods

In this section, we describe the methods used to preprocess the data, implement the Perceptron model, and evaluate its performance on the diabetes prediction task.

3.1. Data Preprocessing

The dataset used in this study is the Pima Indian Diabetes dataset, which consists of 768 instances with 8 features each, representing medical attributes such as glucose levels, body mass index (BMI), and insulin. The target variable is binary, indicating whether the patient has diabetes (1) or not (0). The raw dataset contains some missing or zero values in important attributes such as glucose, insulin, and BMI.

To handle missing values, the following preprocessing steps were applied:

- **Handling Missing Values:** We replaced zero values in columns such as `Glucose`, `BloodPressure`, `SkinThickness`, `Insulin`, and `BMI` with NaN values. These were then imputed by the mean value of the respective columns.

- **Feature Engineering:** A new feature, `Glucose_to_Insulin_Ratio`, was introduced to capture the relationship between glucose and insulin levels. This ratio is calculated by dividing glucose by insulin plus 1 to avoid division by zero.

- **BMI Categories:** The continuous BMI values were binned into four categories (Underweight, Normal, Overweight, Obese), which were subsequently one-hot encoded to capture categorical information.

- **Normalization:** The numerical features were scaled using the `StandardScaler`, which transforms the data to have zero mean and unit variance, making the model more sensitive to feature relationships. The features normalized include `Pregnancies`, `Glucose`, `BloodPressure`, `SkinThickness`, `Insulin`, `BMI`, `Age`, `DiabetesPedigreeFunction`, and the engineered `Glucose_to_Insulin_Ratio`.

3.2. Perceptron Model

The Perceptron algorithm, introduced by Rosenblatt [7], is a linear classifier designed for binary classification tasks. In our study, we implemented a multi-layer Perceptron (MLP) with two hidden layers, which extends the original Perceptron by introducing non-linear activation functions to model more complex relationships between features and the target variable. The model architecture is as follows:

- **Input Layer:** The input consists of 13 features (8 original features + 1 engineered feature + 4 one-hot encoded BMI categories).

- **Hidden Layer 1:** The first hidden layer consists of 72 neurons with ReLU activation. This layer captures non-linear relationships between the input features.

- **Hidden Layer 2:** The second hidden layer consists of 64 neurons, also with ReLU activation, further refining feature representations.

- **Output Layer:** The output layer consists of a single neuron, which uses a sigmoid activation function to output a probability value between 0 and 1, representing the likelihood of a patient having diabetes.

3.3. Training Procedure

The model was trained using the following procedure:

- **Loss Function:** Binary cross-entropy was used as the loss function to optimize the model for binary classification.

- **Optimizer:** Adam optimizer [2] was employed with a learning rate of 0.001. Adam combines the advantages of both adaptive learning rates and momentum to accelerate the training process.

- **Batch Size:** The entire dataset was used in a single batch (batch gradient descent) during each epoch.

- **Epochs:** The model was trained for 40 epochs, during which the model's weights were updated after each epoch based on the loss gradient.

During training, the model weights were initialized randomly but controlled by setting a fixed seed for reproducibility. To avoid overfitting, the training was stopped after 40 epochs, and early stopping was not applied as the performance was evaluated using the test set.

3.4. Evaluation Metrics

To evaluate the performance of the Perceptron model, we employed several standard metrics for binary classification:

- **Accuracy:** The overall correctness of the model's predictions, calculated as the ratio of correctly predicted instances to the total instances.

- **Receiver Operating Characteristic (ROC) Curve and AUC:** The ROC curve plots the true positive rate against the false positive rate at various threshold levels. The area under the ROC curve (AUC) was computed to assess the model's discriminatory power.

We also utilized the `classification_report` from Scikit-learn to compute precision, recall, and F1-score, providing a comprehensive understanding of the model's predictive performance.

4. Experimental Analysis

In this section, we present the experimental results of applying the multi-layer Perceptron model to the Pima Indian Diabetes dataset.

4.1. Training Loss Curve

The model was trained for 40 epochs, and the loss was recorded at each epoch to monitor the training process. Figure 1 shows the training loss curve, which indicates that the loss steadily decreased from approximately 0.70 to 0.50 as the training progressed. This decline suggests that the model is learning and converging towards a solution. The smooth downward trend also suggests that the model is not overfitting within the training epochs.

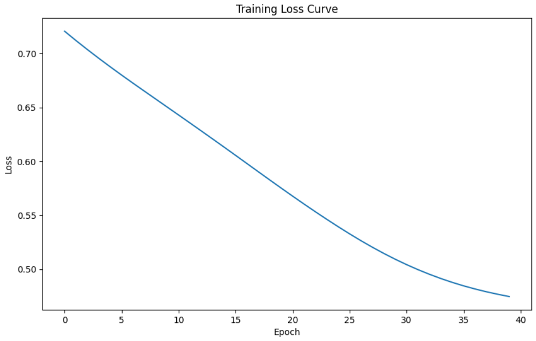

Figure 1. Training Loss Curve over 40 Epochs

4.2. Confusion Matrix

The confusion matrix (Figure 2) shows the distribution of predicted labels versus the true labels on the test set. The matrix highlights 94 true negatives and 35 true positives, with 11 false positives and 14 false negatives. This breakdown gives insight into the types of errors the model is making, especially in an imbalanced dataset like this one. Overall, the confusion matrix demonstrates a reasonable balance between correctly predicted positives and negatives.

4.3. ROC Curve and AUC

The receiver operating characteristic (ROC) curve and the area under the curve (AUC) score are used to measure the model's ability to discriminate between the diabetic and non-diabetic classes. Figure 3 shows the ROC curve, and the AUC score is **0.90**, which provides a summary measure of the model's discriminative power. A higher AUC score indicates better performance in distinguishing between the two classes.

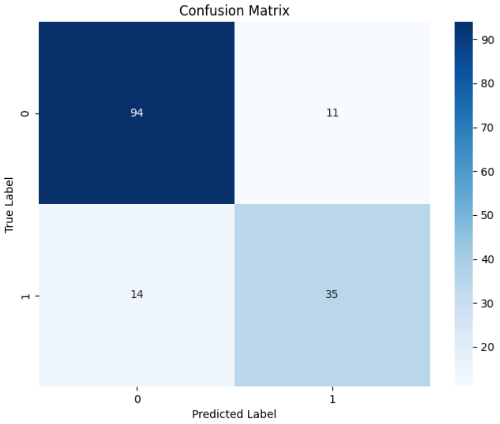

Figure 2. Confusion Matrix for Perceptron Model on Test Set

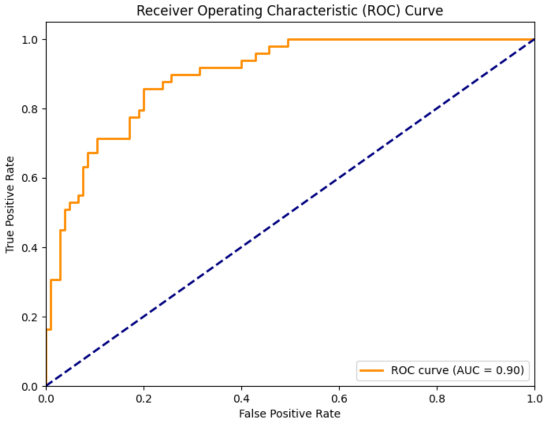

Figure 3. ROC Curve for Perceptron Model (AUC = 0.90)

4.4. Performance Metrics

The class distribution was as follows: 500 instances of non-diabetics and 268 instances of diabetics. The model achieved an overall accuracy of 83.77%. The detailed classification report is shown below:

- Precision for non-diabetics (class 0): 0.87
- Precision for diabetics (class 1): 0.76
- Recall for non-diabetics (class 0): 0.90
- Recall for diabetics (class 1): 0.71
- F1-score for non-diabetics (class 0): 0.88
- F1-score for diabetics (class 1): 0.74

The accuracy on the test set was 83.77%. This performance demonstrates the model's effectiveness at predicting diabetes.

5. Hyperparameter Search

In order to optimize the performance of the model, we conducted a hyperparameter search focusing on the learning rate. Various values of the learning rate were tested, and the corresponding accuracy results are summarized in Table 1.

Table 1. Accuracy results for different learning rates.

Learning Rate	Accuracy (%)
0.0001	40.26
0.0005	77.92
0.001	83.77
0.002	80.52
0.005	81.17
0.01	80.52
0.02	81.17
0.05	79.22
0.1	77.92
0.2	69.48

As shown in Table 1, the best accuracy of 83.77% was achieved with a learning rate of 0.001. The accuracy performance begins to degrade for learning rates higher than 0.005, indicating the importance of fine-tuning this hyperparameter. A graphical representation of the learning rate search is provided in Figure 4.

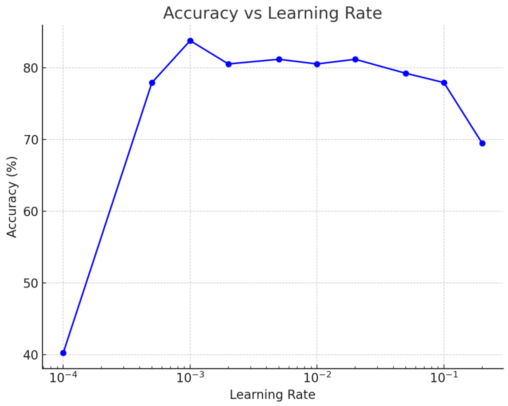

Figure 4. Accuracy as a function of the learning rate. The learning rate was varied on a logarithmic scale, with the highest performance observed at 0.001.

This hyperparameter search confirms that small learning rates, particularly around 0.001, provide the best trade-off between convergence speed and model accuracy for this task.

5.1. Discussion

The experimental results demonstrate that the Perceptron model performs reasonably well in predicting diabetes based on the available features. The loss curve shows a steady decrease, indicating successful convergence during training. The confusion matrix reveals a balanced performance between the two classes, though there are some misclassifications. The ROC curve, combined with the AUC score of 0.90, suggests that the model has good discriminative power. Future improvements could include fine-tuning hyperparameters and exploring more complex neural network architectures to further boost performance.

6. Code Overview

The code for this project is implemented in Python using popular deep learning libraries such as sklearn and pytorch.

You can find the full implementation at the following GitHub repository: `https://github.com/yyaaoo33/deeplearning-ass1.git`.

7. Conclusion

In this study, we developed a multi-layer Perceptron (MLP) model to predict diabetes using the Pima Indian Diabetes dataset. After preprocessing the data, the MLP achieved reasonable accuracy, leveraging non-linear relationships in medical features like glucose levels and BMI. However, class imbalance led to challenges in correctly identifying diabetic patients, as reflected by some false negatives which could be mitigated in future works.

References

[1] A. Begum, S. A. David, D. Hemalatha, M. C. M. Belinda, N. R. Naveena, and V. Cyrilraj. Machine learning approaches to diabetes prediction: A comprehensive evaluation. In *Artificial Intelligence and Information Technologies*, pages 137–143. CRC Press, 2024.

[2] P. K. Diederik. Adam: A method for stochastic optimization. *(No Title)*, 2014.

[3] N. S. El_Jerjawi and S. S. Abu-Naser. Diabetes prediction using artificial neural network. 2018.

[4] R. Ghorbani and R. Ghousi. Predictive data mining approaches in medical diagnosis: A review of some diseases prediction. *International Journal of Data and Network Science*, 3(2):47–70, 2019.

[5] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[6] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant. *Applied logistic regression*. John Wiley & Sons, 2013.

[7] F. Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

[8] M. Soni and S. Varma. Diabetes prediction using machine learning techniques. *International Journal of Engineering Research & Technology (IJERT)*, 9(09):2278–0181, 2020.