# Predict Diabetes using Perceptron

PRIYA PATEL
A1876129

## Abstract

*This study delves into the vital realm of diabetes prediction, employing the Multilayer Perceptron (MLP) algorithm for accurate and early identification. Through a detailed method description and experimental analysis, our research highlights the MLP's proficiency in classifying diabetes cases. While demonstrating notable learning capabilities, the model's performance indicates areas for improvement, particularly in minimizing false positives and negatives. This research contributes to the ongoing efforts in healthcare, emphasizing the significance of advanced machine learning techniques in addressing global health challenges.*

## 1. Introduction

This research focuses on the global issue of diabetes, a widespread chronic condition affecting millions of people. Early and precise identification of diabetes is vital for effective medical care and control. The issue we are addressing involves forecasting the presence of diabetes in individuals based on a range of health-related factors. The dataset employed in this study includes crucial variables like pregnancies, glucose levels, blood pressure, skin thickness, insulin levels, BMI (Body Mass Index), diabetes pedigree function, and age. The primary objective is to precisely identify whether a person has diabetes (1) or does not have diabetes (0) using these specific attributes.

To address this challenge, we utilize the Multilayer Perceptron (MLP) algorithm, which is a form of artificial neural network recognized for its capacity to represent intricate associations in data. The MLP is composed of numerous layers of interconnected neurons, comprising an input layer, one or more hidden layers, and an output layer. Each neuron applies a nonlinear activation function, enabling the model to capture intricate patterns and interrelationships in the data. The goal is to create a predictive model capable of categorizing individuals as either diabetic or non-diabetic, relying on pertinent health indicators.

## 2. Method Description

The Multilayer Perceptron (MLP) is a versatile artificial neural network used for tasks like classification and regression. It consists of layers of interconnected neurons, including input, hidden, and output layers. Neurons use activation functions like ReLU or Sigmoid to introduce non-linearity. The network is trained using backpropagation, adjusting weights to minimize prediction errors through gradient descent optimization.

**Confusion Matrix:** The Confusion Matrix is a crucial tool for assessing classification model performance, such as the Multilayer Perceptron (MLP). It breaks down predictions into four categories: True Positives (correctly identified positives), True Negatives (correctly identified negatives), False Positives (incorrectly identified positives), and False Negatives (missed positives). These metrics help compute essential measures like accuracy, precision, recall, specificity, and the F1-score, offering a comprehensive view of how well the model performs, particularly in scenarios with imbalanced classes.

**RoC Curve:** The ROC Curve is a graphical representation illustrating how a classification model's performance changes with different decision thresholds. It plots the True Positive Rate (TPR or Sensitivity) against the False Positive Rate (FPR) as these thresholds vary. The ROC Curve helps visualize the balance between accurately identifying positive cases and erroneously categorizing negative cases.
True Positive Rate (TPR): This measures the model's ability to correctly classify actual positive instances (TP / (TP + FN)), also referred to as Sensitivity or Recall.
False Positive Rate (FPR): This gauges the model's tendency to incorrectly label actual negative instances as positive (FP / (FP + TN)).

The area under the ROC Curve (AUC-ROC) quantifies how well the model distinguishes between positive and negative classes. A higher AUC-ROC value (closer to 1) indicates superior discrimination, while 0.5 signifies random guessing.

**Effects and Deficiencies:**
Confusion Matrix: The Confusion Matrix provides a detailed breakdown of a model's performance, allowing us to understand its strengths and weaknesses. For example, it helps identify scenarios where the model has high precision but low recall, or vice versa. It highlights the impact of false positives and false negatives on the overall evaluation.
ROC Curve: The ROC Curve visualizes a model's discrimination ability across various thresholds. It illustrates how the TPR and FPR change as the decision boundary shifts. However, it does not provide a single threshold-independent metric like accuracy. Additionally, in cases of imbalanced datasets, it may not fully represent the model's performance, and alternative metrics like the Precision-Recall Curve can be more informative.

## 3. Experimental Analysis

In this study, I conducted a comprehensive experimental analysis to evaluate the performance of the model using the confusion matrix and accuracy metrics. The primary aim was to understand how well the model performs on both the training and testing data, providing insights into its generalization capabilities.

**Confusion Matrix Evaluation:**
The confusion matrix is a fundamental tool in binary classification tasks. It provides detailed information about the model's performance, breaking down predictions into true positives, true negatives, false positives, and false negatives.
Training Accuracy:
The training accuracy is a measure of how well the model fits the training data. It indicates the percentage of correct predictions on the training set.
Testing Accuracy:
The testing accuracy evaluates the model's performance on unseen data, providing an estimate of its ability to generalize to new, unseen instances.

**Confusion Matrix Analysis:**
The confusion matrix was computed to understand the distribution of predictions on both the training and testing sets. This allowed for a detailed examination of the model's strengths and weaknesses.

**Training vs. Testing Accuracy:**
The training and testing accuracies were calculated to assess the model's performance on both seen and unseen data. This helps in identifying potential overfitting or underfitting issues.

**Confusion Matrix Findings:**

The confusion matrix revealed that the model had a relatively balanced performance between true positives and true negatives. However, there were notable instances of false positives and false negatives. This suggests that the model may benefit from further refinement.
**Training vs. Testing Accuracy:**
The training accuracy of 72.96% indicates that the model learned reasonably well from the training data. However, the testing accuracy of 62.99% suggests that the model's generalization capabilities may have room for improvement.

## 4. Code

The code link is:
https://github.com/Priyapatel1204/Deep_Learning_Assignment/blob/main/Assignment1_DLF.ipynb

## 5. Conclusion

In conclusion, this research addresses the pressing global issue of diabetes prediction using the Multilayer Perceptron (MLP) algorithm. Our study underscores the significance of early and accurate diagnosis for effective healthcare management. The experimental analysis revealed that while the MLP exhibited a commendable performance in correctly classifying diabetes cases, there remains potential for refinement, especially in minimizing false positives and false negatives. While the model displayed a strong learning capacity, as indicated by the training accuracy, further work is needed to enhance its generalization capabilities, as reflected in the testing accuracy. This research contributes to the ongoing pursuit of precise diabetes prediction, with the ultimate aim of improving healthcare outcomes for millions worldwide.

## References


[1] https://machinelearningmastery.com/how-to-develop-multilayer-perceptron-models-for-time-series-forecasting/
[2] https://www.kaggle.com/code/palmer0/predicting-diabetes-with-multilayer-perceptrons
[3] https://www.kaggle.com/code/vipullrathod/binary-classification-with-multilayer-perceptrons
[4] https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-imbalanced-classification/#:~:text=The%20simplest%20confusion%20matrix%20is,positive%20(class%201)%20classes.&text=The%20metrics%20that%20make%20up,cells%20in%20the%20confusion%20matrix.