Deep Learning Analysis on Pima Diabetes Dataset

VuQuoc An - Adelaide University

Abstract

*Diabetes, a chronic ailment affecting millions worldwide, necessitates precise diagnostic tools. The Pima Indian heritage dataset, sourced from the National Institute of Diabetes and Digestive and Kidney Diseases, offers a platform to harness deep learning for diagnostic predictions. This research employs a Multilayer Perceptron (**MLP**) to develop a predictive model and delves deep into the exploration of various hyperparameters to understand their impact on the model's performance. Intriguingly, due to the linearity of the data, hyperparameter tuning revealed that the model's performance was comparably robust, if not superior, even when reduced to a single perceptron. This finding underscores that complex architectures might not always yield better results, especially when data exhibits strong linear characteristics. By systematically tuning hyperparameters and emphasizing data preprocessing, this study aims to optimize the diagnostic capabilities of the **MLP** model, highlighting the potential of machine learning in medical diagnostics.*

1. Introduction

Diabetes mellitus, characterized by chronic hyperglycemia, is a global health challenge. Accurate and timely diagnosis is paramount for effective intervention [6]. The Pima Diabetes dataset, encompassing medical diagnostic metrics, offers a unique perspective on diabetes predictors within the Pima Indian heritage female population.

Deep learning, with its prowess in pattern recognition, holds promise in medical diagnostics. This research leverages the Pima Diabetes dataset to explore the impact of various hyperparameters on the performance of a Multilayer Perceptron (**MLP**). The study underscores the importance of data preprocessing, especially median imputation for handling missing values, and systematically investigates the influence of different hyperparameters, such as the number of layers, neurons, regularization techniques, and activation functions. Through this comprehensive approach, the research aims to provide insights into the optimal configuration of deep learning models for diabetes diagnosis.

2. Related Research

In clinical diagnosis problems, classification plays a vital role in further treatment of the disease. Various studies have been done on the diabetes data classification using Pima Indian diabetes dataset [1] [2] [3] [5]. Using the dataset from University of California, Irvine (UCI) machine many methods for classification have been develop and enhanced.

Kannadasan et al. (2019) proposed a deep learning framework for diabetes data classification using stacked autoencoders. Their approach achieved a classification accuracy of 86.26%, emphasizing the potential of deep learning techniques in medical diagnostics [3].

A study in 2020, using the Pima Indian Diabetes dataset, employed machine learning techniques such as logistic regression, decision tree, random forest, and gradient boosting machine. Intriguingly, their proposed model reported an exceptionally high accuracy rate of 98% [5]. Such a rate eclipses most other reported accuracies and raises questions about potential overfitting or the data leakage issues.

Butt et al. (2021) explored various machine learning algorithms for diabetes classification and prediction. Their study highlighted the significance of feature selection in enhancing predictive model performance, achieving an accuracy of 78.5% with the random forest algorithm [2].

Chang et al. (2022) utilized a range of machine learning models, including logistic regression, decision trees, and random forests, for predicting Type 2 diabetes. Their research emphasized the importance of feature selection, with the random forest algorithm achieving an accuracy of 78.5% [1].

3. Methodology

3.1. Data Preprocessing

The Pima Indian Diabetes dataset, sourced from Kaggle [4], consists of 768 entries spanning nine columns, each detailing various health metrics of the subjects. While the dataset is free from null or missing values, it does exhibit certain inconsistencies. Specifically, attributes such as glucose concentration (Gluc), blood pressure (BP), skin fold thickness (Skin), insulin, and BMI present zero values. Based on domain knowledge, these zero values fall outside the normal range and are therefore considered inaccurate.

Data preprocessing plays a pivotal role in determining the performance of machine learning models, especially for datasets like the Pima Indian Diabetes dataset, which exhibits linearity and has a relatively low dimensionality. Consequently, the first preprocessing step was to impute these invalid zero values with the median of the respective features. This median imputation ensures that the data distribution remains unskewed, offering a more accurate representation of the underlying health metrics.

Given the dataset's modest size, efforts were made to retain as many rows as possible, avoiding unnecessary data loss. After imputation, due to the linear nature of the data and the use of the **MLP** model, the data was scaled using the min-max scaling technique.

3.2. Model Architecture

3.2.1 Multilayer Perceptron (MLP)

The primary model employed in this study is the multi-layer perceptron (**MLP**). An **MLP** is a feedforward neural network that consists of multiple layers of neurons. The output y of a neuron is given by:

$$y = f(\mathbf{w} \cdot \mathbf{x} + b)$$

Where:

- $\mathbf{x}$ represents the input vector.
- $\mathbf{w}$ denotes the weight vector.
- b is the bias.
- f is the activation function, such as sigmoid or ReLU.

For layers with multiple neurons, the output can be represented as:

$$\mathbf{Y} = f(\mathbf{W}\mathbf{X} + \mathbf{b})$$

Where f is applied element-wise, and $\mathbf{W}$ and $\mathbf{b}$ are the weight matrix and bias vector, respectively.

To determine the optimal configuration for the **MLP**, a systematic grid search was conducted, exploring various hyperparameters, including the number of layers, neurons per layer, activation functions, and regularization techniques.

Grid search, while comprehensive, can be computationally demanding and may lead to overfitting. To address these challenges, this study adopted a chained grid search approach, tuning hyperparameters sequentially. This method not only economizes computational resources but also narrows the search space, potentially reducing overfitting risks.

However, this approach assumes hyperparameters are independent, which might not always be the case. For instance, the optimal value for one hyperparameter might be influenced by another. There's also the risk of settling for a local optimum, missing out on a globally optimal configuration. The study acknowledges these potential pitfalls, ensuring a balanced interpretation of the results.

The Multilayer Perceptron (**MLP**) is a type of feedforward artificial neural network. It comprises multiple layers of nodes, including an input layer, hidden layers, and an output layer. Each node in one layer connects to every node in the subsequent layer with associated weights.

3.3. Hyperparameter Tuning via GridSearch

The hyperparameter tuning process is structured in a sequential manner using GridSearch. The search order is as follows:

1. Optimize the number of neurons and layers.
2. Fine-tune L1 and L2 regularization parameters.
3. Determine the optimal dropout rate.
4. Identify the best activation function.

The specific hyperparameters under consideration are:

Neurons in Hidden Layers: [1, 2, 4, 8, 16, 32]

Number of Layers: [0, 1, 2, 3, 4]

L1 Regularization: [0, 0.001, 0.01, 0.1]

L2 Regularization: [0, 0.001, 0.01, 0.1]

Dropout Rate: [0, 0.1, 0.2, 0.3, 0.4]

Activate Function: [Logistic, tanh, Relu, Leaky ReLU]

3.4. Model Evaluation

3.4.1 Data Splitting

For theexperiments, the Pima Indian Diabetes dataset was divided into training, validation, and test sets using an 8:1:1 ratio. This means that 80% of the data was used for training the model, 10% was used for validation during the training process to tune hyperparameters and prevent overfitting, and the remaining 10% was reserved for testing the model's performance on unseen data.

Given the imbalanced nature of our dataset, with 268 positive (diabetic) instances out of a total of 768, it's crucial to ensure that the distribution of the classes is consistent across the training, validation, and test sets. To achieve this,stratified sampling is employed during the data splitting process. Stratification ensures that each split has a similar proportion of positive (diabetic) and negative (non-diabetic) instances, preserving the original distribution of the dataset. This method is particularly beneficial as it ensures that the model is exposed to a representative sample of both classes during training and evaluation, mitigating potential biases and providing a more accurate assessment of the model's performance.

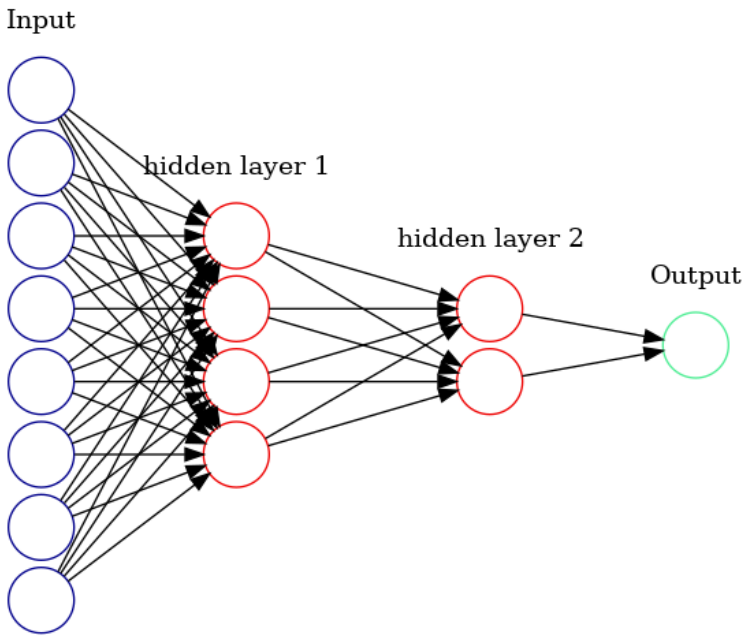

Figure 1. Architecture of a typical MLP model.

3.4.2 Evaluation Metrics

In the context of the research, which focuses on binary classification (diabetic or non-diabetic), accuracy is a primary metric. Accuracy is defined as the ratio of correctly predicted instances to the total instances in the dataset. Mathematically, it can be represented as:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \quad (1)$$

4. Experimental and Results

4.1. Experimental Setup

4.1.1 Number of Neurons and Layers

The experiments varied the number of neurons: [1, 2, 4, 8, 16, 32] and the number of layers: [0, 1, 2, 3, 4]. Interestingly, when the model has zero layers, it simplifies to a perceptron, analogous to a logistic regression model. The results suggest that the optimal configuration hovers between 0 and 1 layer, pointing towards a linear relationship in the data.

A notable trend in the accuracy graph (Figure 2) is the decline in performance towards its bottom left. This decline is symptomatic of overfitting, where an overly complex model begins to capture noise rather than genuine data patterns.

Furthermore, when the model contains only one neuron per layer, adding more layers results in decreased accuracy. This phenomenon can be attributed to the dilution of information across layers. In essence, the model transforms into a series of single neurons, each transmitting an increasingly distorted version of the input, without enhancing its representational capacity.

To further validate the linear nature of the data, we can consider the efficacy of Linear Discriminant Analysis (LDA) on similar datasets. LDA, which thrives on linearly separable data, aims to maximize the inter-class mean distance while minimizing intra-class scatter. Saxena's work [7] with LDA on the Pima Diabetes dataset achieved an accuracy of 78.5%, underscoring the dataset's linear characteristics. This linearity is also supported by the intuitive understanding that certain medical metrics, such as body weight or age, often have linear correlations with conditions like diabetes.

4.1.2 L1 and L2 Regularization

Having established the model architecture (a single perceptron), we proceeded to apply L1 and L2 regularization. The results, as depicted in the figure 3,suggest that regularization techniques have limited impact on this dataset. This is likely due to the dataset's simplicity, comprising only 8 attributes.

Regularization techniques like L1 and L2 are typically

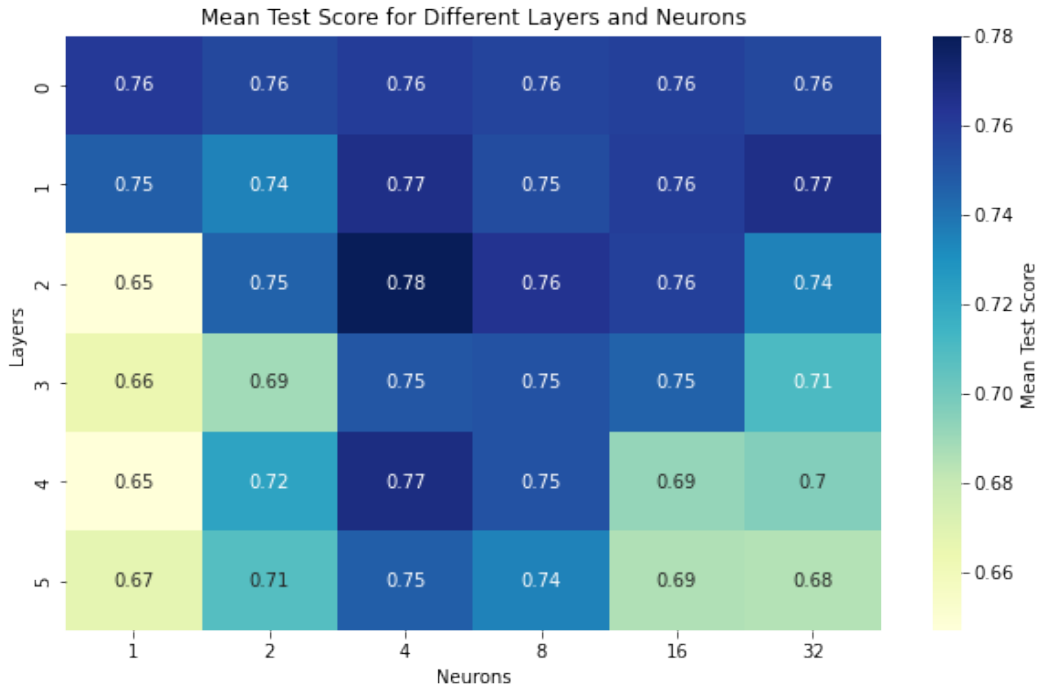

Figure 2. Accuracy of models by number of layers and neurons

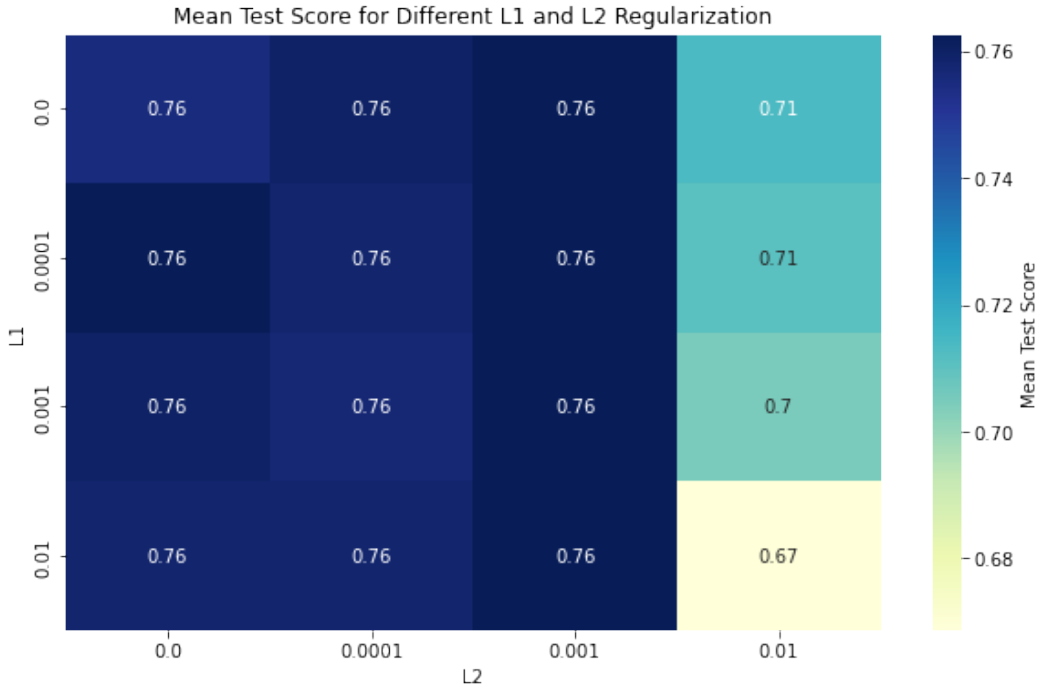

Figure 3. Accuracy of models with varying L1 and L2 regularization

more effective for complex models where there's a need to mitigate overfitting. Effectively, the result show that there are no affect of regulazation on the model. Consequently, no L1 and L2 regularization are selected as the optimal hyperparameters.

4.1.3 Dropout

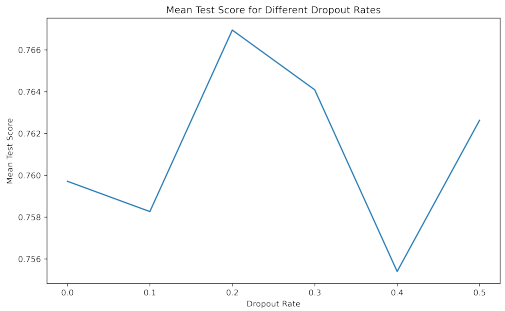

Figure 4. Accuracy variations by dropout rate

The dropout results, depicted in Figure 4, show minimal variation and lack a discernible trend, suggesting that dropout doesn't significantly influence this simple dataset. Consequently, no dropout was selected as the optimal hyperparameter setting.

4.1.4 Activation Function

Considering the model's architecture, which essentially reduces to logistic regression, the choice of activation function becomes irrelevant.

4.1.5 Final Result

Following the hyperparameter optimization, the optimal model parameters were identified as:

- No L1 or L2 regularization
- No dropout
- Singular perceptron configuration

The performance metrics for this configuration are:

F1 Score: 0.6667
Recall: 0.625
Accuracy: 0.8052

4.2. Discussion

The experimental findings shed light on the nuances of the Pima Diabetes dataset and the potential of deep learning models, with a specific emphasis on the **MLP**. One of the most striking outcomes was the realization that a single perceptron could rival, or even outperform, more intricate architectures. This underscores the pivotal role of understanding the data's intrinsic characteristics before diving into complex model designs. The dataset's inherent linearity, as highlighted by the robust performance of the single perceptron, advocates for the efficacy of simpler models in certain scenarios. This aligns with the tenets of Occam's razor, which promotes the idea that simplicity often trumps complexity when seeking solutions. Moreover, the pivotal role of data preprocessing, especially the strategy of median imputation, emerged as a cornerstone for model training, ensuring that the model learns from a genuine representation of the data, thereby bolstering its predictive reliability.

4.3. Limitations and Future Work

Despite the insights gleaned from the study, certain limitations warrant mention. The Pima Diabetes dataset, while informative, is constrained in size, potentially affecting the broader applicability of the conclusions. Moreover, the dataset's focus on a niche demographic, namely females of Pima Indian descent, raises questions about its representativeness for a more diverse population.

Looking ahead, the modest accuracy achieved points towards the potential of refining the data handling process. Given the indications that the dataset lacks intricate underlying patterns, efforts could be channeled towards data cleansing or enrichment. Incorporating additional data or leveraging data augmentation strategies might pave the way for enhanced model efficacy. Furthermore, the exploration of unsupervised learning paradigms, such as clustering or dimensionality reduction, could be instrumental in unveiling latent patterns or intricate relationships within the dataset.

5. Code

The code of the research could be find at: https://github.com/vuquocan1987/Assignment1DeepLearning.git

6. Conclusion

This research delved into the Pima Diabetes dataset, employing a Multilayer Perceptron (**MLP**) to predict diabetes onset. Through systematic hyperparameter tuning and data preprocessing, we discovered the surprising efficacy of a single perceptron, challenging the notion that complex architectures always yield better results. thefindings underscore the importance of understanding the data's nature and the potential of simpler models in specific scenarios. As machine learning continues to make strides in medical diagnostics, it's imperative to approach problems with a nuanced understanding, ensuring that the chosen solutions are not only effective but also efficient.

References

[1] UM Butt, S Letchmunan, M Ali, FH Hassan, A Baqir, and HHR Sherazi. Machine learning based diabetes classification and prediction for healthcare applications. *J Healthc Eng*, 2021:9930985, 2021. 1

[2] V Chang, J Bailey, QA Xu, and Z Sun. Pima indians diabetes mellitus classification based on machine learning (ml) algorithms. *Neural Computing and Applications*, pages 1–17, 2022. Epub ahead of print. 1

[3] K Kannadasan, Damodar Reddy Edla, and Venkatanareshbabu Kuppili. Type 2 diabetes data classification using stacked autoencoders in deep neural networks. *Clinical Epidemiology and Global Health*, 7(4):530–535, 2019. 1

[4] UCI Machine Learning. Pima indians diabetes database, 2019. 1

[5] H Naz and S Ahuja. Deep learning approach for diabetes prediction using pima indian dataset. *J Diabetes Metab Disord*, 19(1):391–403, 2020. 1

[6] World Health Organization. Diabetes: Key facts, 2022. Accessed: [Insert the date you accessed the website]. 1

[7] Arnav Saxena. Lda 78.5% for indian pima diabetes, 2023. 3