A Perceptron-Based Approach for Predicting Diabetes

Nithin Shankar Chinnasamy
The University of Adelaide
nithin.shankar@student.adelaide.edu.au

Abstract

Diabetes is also among the most common causes of morbidity and mortality in the population of the Earth. The diagnosis of diabetes is fundamental in the management of diabetes. In this study, we propose a categorization approach for diabetes data using perceptron in view of the enhancement in neural networks for improved accuracy and efficiency. The dataset is composed of 768 women aged more than 21 years. I used this model for training and tested it at varying learning rate. This results an accuracy rate of 80.8% with an optimistic learner rate of 0.05.[7]

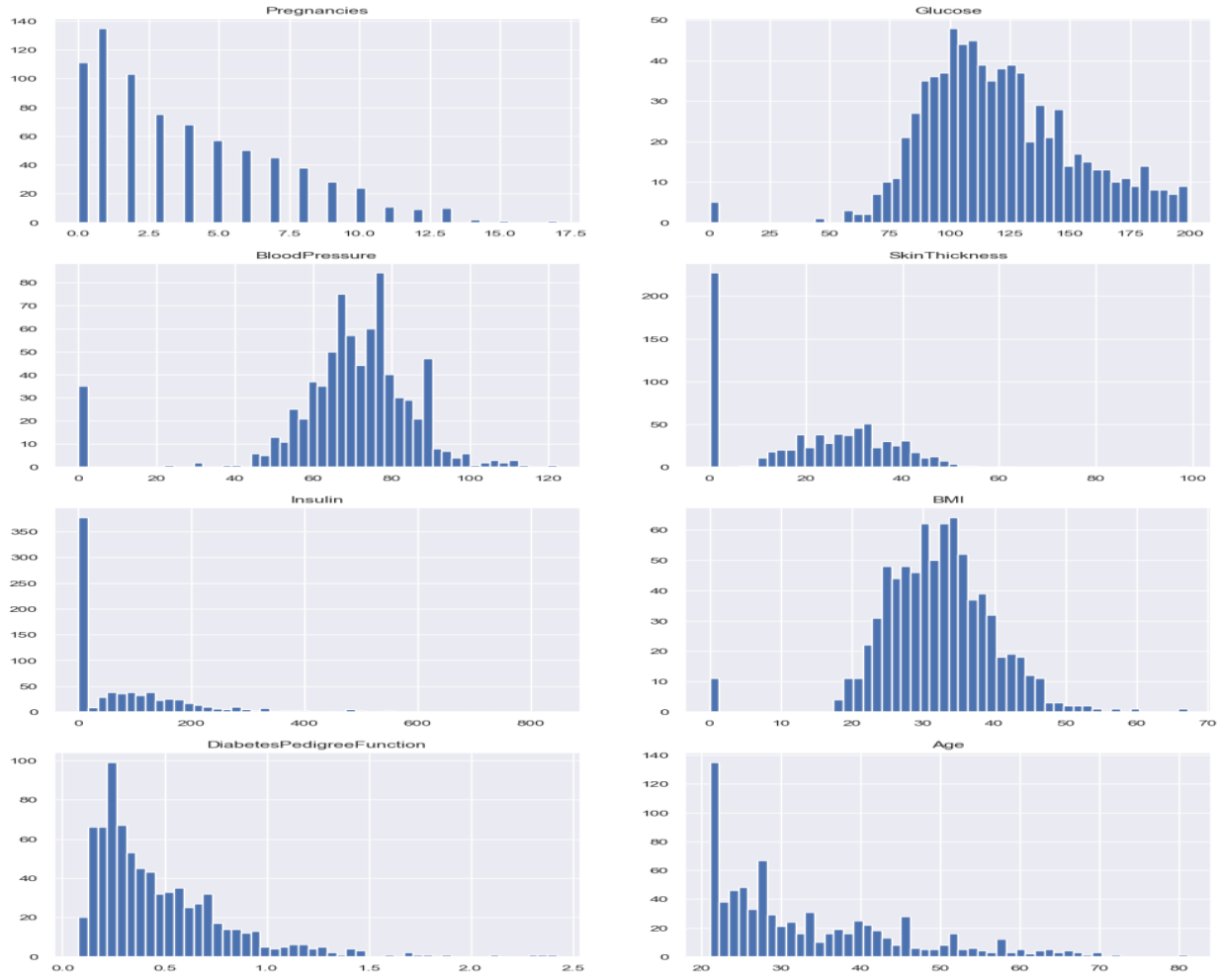


Fig 1. Histogram for each attribute

1. Introduction

Diabetes is one of the most common chronic diseases that impact health in the world. The use of machine learning in the accurate and timely prediction of this condition is therefore very important to the management and treatment process. In this work, we examine the use of a Perceptron model, which is part of a neural network, in diagnosing diabetes through medical records. The dataset used is the Pima Indians Diabetes Dataset, which majority of it consists of eight attributes including blood glucose level, BMI and age. [14]As a single-layer neural network, the Perceptron aggregates input features linearly and, employing an activation regulation, predicts the diabetic status of a patient.

The primary goal of this study is to assess the performance of a Perceptron for this binary classification task, particularly focusing on the effect of different learning rates. The model's effectiveness will be evaluated using accuracy, AUC scores, and execution time.

---

2. Method

2.1. Dataset Description

The Pima Indians Diabetes Dataset is the common dataset used for binary classification levels in machine learning. It has 768 examples of them; each case refers to a female Indian Pima patient. The features extracted from the dataset include glucose concentration, insulin level, BMI, and age with the target as dichotomous- diabetes or no diabetes. To prepare the features for the subsequent layers, normalization was performed to improve the speed at which the model is trained.[17]

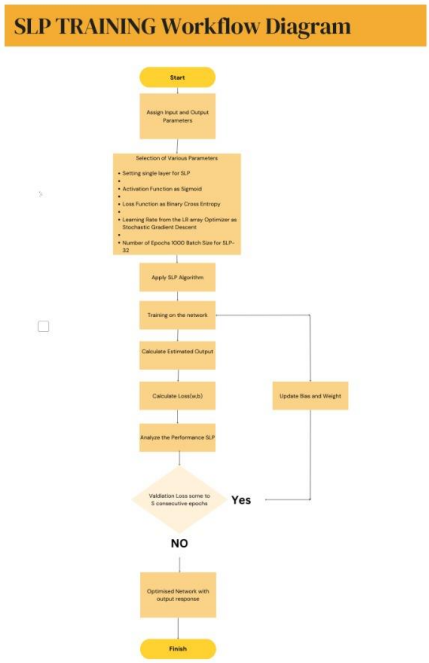
2.2. Perceptron Model

A Perceptron is a basic form of a neural network that takes input values, weighted by weights, passes them through an activation function, and using the total to make a prediction. In this case we used the sigmoid activation function which gives us a probability of the output for each class. This allows the Perceptron to serve as a binary classifier – that is; to classify how likely it is for a particular patient to have diabetes. In this sense, the model's weights were updated according to SGD, an algorithm that directly modifies the weights to reduce the rate of classification error endorsed by the learning rate. Experiments with learning rate and influence on the performance of the model were conducted with a set of values for learning rate.[24]

Weight Update Equation:

$$\boldsymbol{Wnew = Wold + \eta(y - y\hat{}\,)x}$$

where η is the learning rate, y is the true label, y^ is the predicted value, and x is the feature vector.



Flow chart

2.3. Training Procedure

The present study utilized stratified 5-fold cross-validation to make sure that the proportion of the diabetes and non-diabetes class in the training and testing sets was maintained. This method splits the data into five sets and the model is trained and tested on all the sets accruing to achieving a comprehensive assessment of the model performance.[28]

Standardization was performed to ensure the normalization with a mean of 0 and standard deviation of 1 for every feature. With the attempt to combat overfitting, various learning rates including from 0.0005 up to 0.3 and early stopping were used; if validation accuracy no longer improved then training was stopped.

---

3. Experimental Analysis

3.1. Performance Metrics

To evaluate the model, we recorded three key metrics:

1. **Accuracy** – The proportion of correct predictions made by the model.
2. **AUC Score** – The area under the receiver operating characteristic (ROC) curve, which measures the ability of the model to discriminate between classes.
3. **Execution Time** – The time taken to train the model across each fold.

For each learning rate, the model was tested using 5-fold cross-validation, and the average of the metrics across the folds was calculated. The results are shown below.

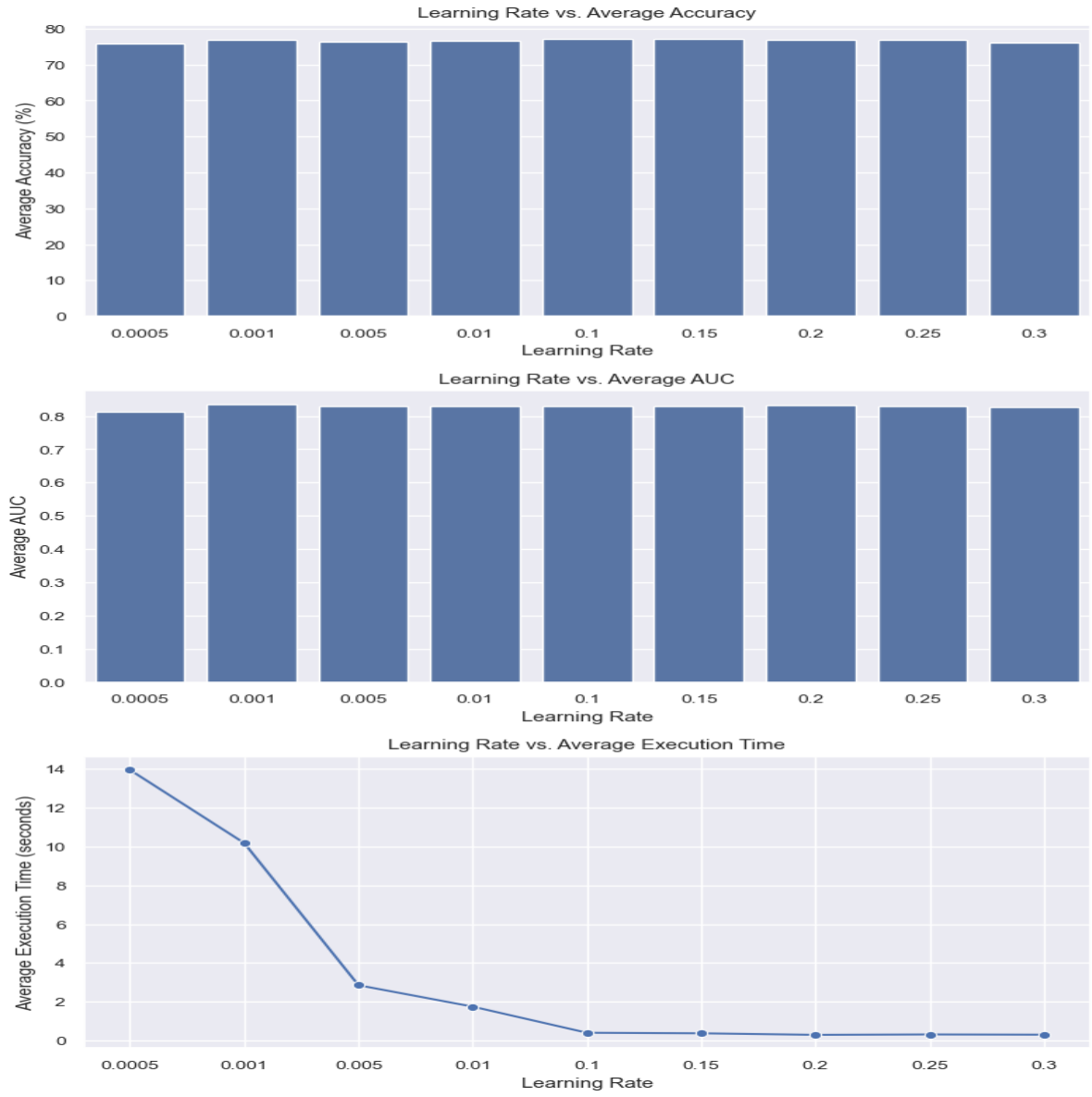


Fig 2. Mean AUC score and Execution times

3.2. Results

- **Accuracy:** The highest mean accuracy of 80.78% was observed at a learning rate of 0.2, with small variations across folds. This indicates that the model is performing well at identifying both diabetic and non-diabetic patients.
- **AUC Score:** The mean AUC score across learning rates was approximately 0.83, indicating that the model had a strong ability to distinguish between classes.

- **Execution Time:** The lowest execution time was recorded for a learning rate of 0.1, where early stopping limited unnecessary training epochs.

3.3. Visualizations

Three key graphs were plotted to visualize the results:

1. **Loss and Accuracy Curves:** These curves show how both the training and validation losses, as well as accuracies, evolved during training for different learning rates.
2. **ROC Curves:** The ROC curves were plotted for each fold, providing a graphical representation of the model's performance at various decision thresholds.
3. **Learning Rate vs. Metrics:** Bar plots were created to show how accuracy, AUC, and execution time changed as the learning rate varied.

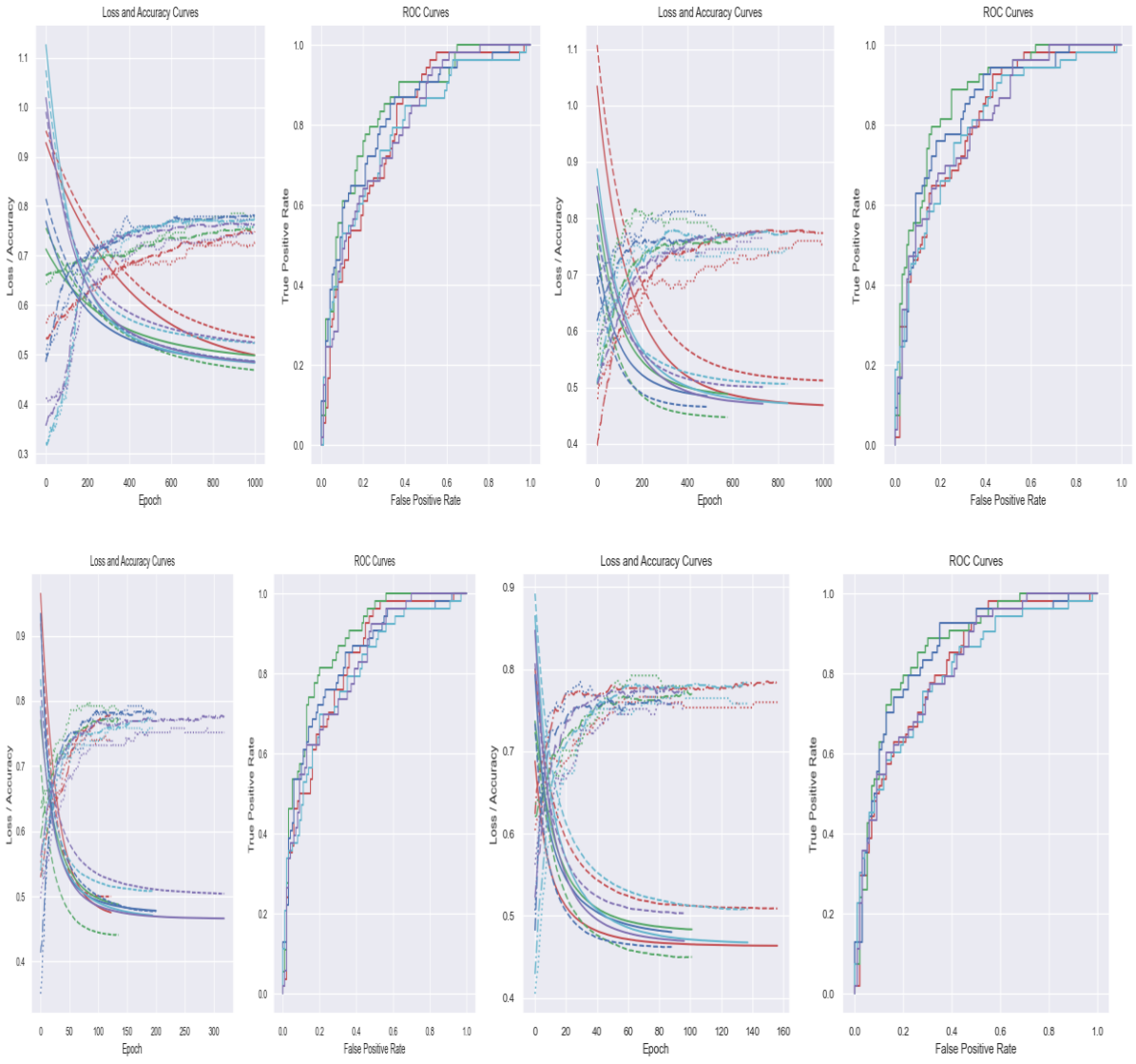


Fig 3. ROC Curves

4. Conclusion

This study successfully trained and assessed a Perceptron model for the purpose of classification of diabetes. Several experiments were conducted with the variation of learning rates, and we determined that the learning rate of 0.2 was the optimal candidate. This was enabled by the early stopping technique which prevented overfitting and led to cutting down on computations that could otherwise have been made.

For the future work, models like multi-layer perceptron or deep neural networks could be used as an attempt to get better classification. Of course, one more step up from here could be hyperparameters tuning and data augmentation for even better results.

5. Code

The github code: https://github.com/nithinamigo/nithinshankar-DiabetesPerceptron. This code contains python libraries.

References

1. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). Available at: CVPR 2020 Author Guidelines.
2. Kaggle: Pima Indians Diabetes Database. Available at: Pima Dataset on Kaggle.
3. Bader Fahad Alkhamees. An optimized single layer perceptron-based approach for cardiotocography data classification. *International journal of advanced computer science applications*, 13(10):239–245, 2022.
4. Franc¸ois Chollet et al. Keras. https://keras.io, 2015.
5. Aman Darolia and Rajender Singh Chhillar. Analyzing three predictive algorithms for diabetes mellitus against the pima indians dataset. *ECS transactions*, 107(1):2697–2704, 2022.
6. Lucas B.V. de Amorim, George D.C. Cavalcanti, and Rafael M.O. Cruz. The choice of scaling technique matters for classification performance. *Applied soft computing*, 133:109924–, 2023.
7. J. J. Forsstrom, K. Irjala, G. Sel¨ en, M. Nystr´ om, and P. Eiu-¨ und. Using data preprocessing and single layer perceptron to analyze laboratory data. *Scandinavian journal of clinical laboratory investigation. Supplement*, 55(S222):75–81, 1995.
8. Tadayoshi Fushiki. Estimation of prediction error by using k-fold cross-validation. *Statistics and computing*, 21(2):137– 146, 2011.

9. Aurelien G´ eron.´ *Hands-on machine learning with ScikitLearn, Keras, and TensorFlow concepts, tools, and techniques to build intelligent systems*. O'Reilly, Sebastopol, CA, third edition. edition, 2023.
10. Mariwan Ahmed Hama Saeed. Diabetes type 2 classification using machine learning algorithms with up-sampling technique. *Journal of Electrical Systems and Information Technology*, 10(1):8–10, 2023.
11. D.J. Hand and C. Anagnostopoulos. When is the area under the receiver operating characteristic curve an appropriate measure of classifier performance? *Pattern recognition letters*, 34(5):492–495, 2013.
12. Charles R. Harris, K. Jarrod Millman, Stefan J. van der Walt,´ Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernandez´ del R´ıo, Mark Wiebe, Pearu Peterson, Pierre Gerard-´ Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with numpy. *Nature (London)*,
585(7825):357–362, 2020.
13. Yi-Chung Hu. Bankruptcy prediction using electre-based single-layer perceptron. *Neurocomputing (Amsterdam)*, 72(13):3150–3157, 2009.
14. John D. Hunter. Matplotlib: A 2d graphics environment. *Computing in science engineering*, 9(3):90–95, 2007.
15. UCI Machine Learning. Pima indians diabetes database, 2016.
16. Basith K. Moien Abdul, Muhammad J. Hashim, Kwan K. Jeffrey, Devi G. Romona, Halla Mustafa, and Juma A. Kaabi. Epidemiology of type 2 diabetes – global burden of disease and forecasted trends. *Journal of Epidemiology and Global Health*, 10(1):107–111, 03 2020. Copyright - © 2020. This work is licensed under http://creativecommons.org/licenses/by-nc/4.0/ (the "License"). Notwithstanding the ProQuest Terms and Conditions, you may use this content in accordance with the terms of the License; Last updated - 2020-12-01.
17. D Mualfah, Y Fatma, and R A Ramadhan. Anti-forensics: the image asymmetry key and single layer perceptron for digital data security. *Journal of Physics: Conference Series*, 1517(1):12106–, 2020.
18. Huma Naz and Sachin Ahuja. Deep learning approach for diabetes prediction using pima indian dataset. *Journal of diabetes and metabolic disorders*, 19(1):391–403, 2020.
19. World Health Organization. *Classification of diabetes mellitus*. World Health Organization, 2019.
20. Radhanath Patra and Bonomali khuntia. Analysis and prediction of pima indian diabetes dataset using sdknn classifier technique. *IOP conference series. Materials Science and Engineering*, 1070(1):12059–, 2021.
21. Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duches-´ nay. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12:2825–2830, 2011.
22. Lutz Prechelt. Automatic early stopping using cross validation: quantifying the criteria. *Neural networks*, 11(4):761– 767, 1998.

23. Rajni Rajni and Amandeep Amandeep. Rb-bayes algorithm for the prediction of diabetic in “pima indian dataset”. *International journal of electrical and computer engineering (Malacca, Malacca)*, 9(6):4866–4872, 2019.
24. F Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386–408, 1958.
25. Kai-Yeung Siu, Amir Dembo, and Thomas Kailath. On the perceptron learning algorithm on data with high precision. *Journal of computer and system sciences*, 48(2):347–356, 1994.
26. Jack W. Smith, J.E. Everhart, W.C. Dickson, W.C. Knowler, and R.S. Johannes. Using the adap learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings - Annual Symposium on Computer Applications in Medical Care*, pages 261–265, 1988.
27. Michael L. Waskom. seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021, 2021.
28. Wes McKinney. Data Structures for Statistical Computing in Python. In Stefan van der Walt and Jarrod Millman, editors,′ *Proceedings of the 9th Python in Science Conference*, pages 56 – 61, 2010.