**ASSIGNMENT 1: PREDICT DIABETES USING PERCEPTRON**

VIDHUN CHUNGATHE VIDHYA

A1903176

**Abstract**

*The aim of this project is to design a machine learning model that will predict diabetes in patients using available demographic and medical data. A few variables included in this dataset are glucose level, BMI, age, among others. Here, all these inputs would predict diabetes using a logistic regression classifier. This model yielded an accuracy of 0.69 on the test dataset. These results point out that major medical factors like glucose level and BMI are strong predictors for diabetes.*

# **Introduction**

Diabetes is a disease that has been found in millions of people around the world. Detecting it as soon as possible is necessary to reduce the burden. Diabetes can be taken care of through various machine learning techniques given with the patient data. In this paper, we have developed a model that can predict diabetes using health parameters such as BMI, Blood pressure, and glucose level. The work in this paper describes about creating a model, predicting the data, evaluating the results of the predicted model.

# **Literature Review**

**I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.**

This book provides an introduction to deep learning. The book gives an introduction to its fundamental concepts in depth, with some contemporary architectures. Authored by leading experts in this field, it stands both as a tutorial and as a reference for students, researchers, and practitioners alike.

**F. Pedregosa, G. Varoquaux, A. Gramfort, et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.**

This paper introduces the scikit-learn library-a widely used open-source Python library for machine learning. In this work, simple and efficient tools for data mining and analysis were built on top of NumPy, SciPy, and matplotlib.

**Kingma, D. P., and Ba, J. Adam: A method for stochastic optimization. Proceedings of the 3rd International Conference for Learning Representations, ICLR 2015.**

The present work introduces the Adam optimizer; nowadays, it is one of the most used optimization algorithms in deep learning.

**H. Zhang, M. Cisse, Y. Dauphin, and D. Lopez-Paz, "mixup: Beyond Empirical Risk Minimization," in Proceedings of the 6th International Conference on Learning Representations (ICLR), 2018.**

Mixup is a data augmentation strategy that departs from conventional empirical risk minimization by mixing pairs of input samples and their labels.

**Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-Based Learning Applied to Document Recognition," Proceedings of the IEEE, vol. 86, no. 11, pp. 2278-2324, Nov. 1998.**

This seminal paper describes how the use of gradient-based learning techniques, specifically convolutional neural networks (CNNs), may be applied to document recognition tasks such as handwriting and character recognition.

**P. J. Werbos, "Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences," Ph.D. dissertation, Harvard University, 1974.**

This is one seminal piece where Paul Werbos first came up with the idea known as back-propagation-a fundamental algorithm in training neural networks.

# **Methodology**

## **Dataset Description and Preprocessing**

The patient data used in this project are comprehensive, ranging from a number of important health indicators usually associated with diabetes. Glucose represents blood sugar for a patient while fasting, body mass index, or BMI, is a way to show body fat based on height and weight, while age is measured in years. The dataset also includes the blood pressure and other important health characteristics such as triceps skin fold thickness, two-hour serum insulin, and the diabetes pedigree function, which is an estimation of diabetes development likelihood determined from genetic and family history. Every patient is labeled as +1 and -1, respectively, for having diabetes or not. This is a much easier label to use for classification tasks, which are the main goals of the current study.

## **Model Selection and Training**

Logistic regression is adopted as the major model in the present binary classification task, considering its simplicity, interpretability, and ability to provide impressive performance in binary outcome predictions. In fact, logistic regression is the best fit when the goal involves segmenting data into two distinct categories, such as identifying whether a person has diabetes or not. The pre-processed dataset used in this model training consisted of 80% training and the remaining 20% testing.

In the training phase, it divided the data into a training set, from which it could learn, and a test set on which it was being tested with data it had never tested. In a logistic regression model, model parameters are tuned in such a way that, guided by the loss measured from the training data, the prediction error can be minimized. After training, various metrics were computed to assess model performance on the test set: accuracy, precision, recall, and F1 score. These measures provide an extended overview of the model balance between precision and recall and its ability to predict diabetic or non-diabetic cases.

## **Evaluation Metrics**

Important measures that are taken into consideration to assess in detail the predictive power of the logistic regression model include the following. Accuracy is a general indicator of performance, as it measures the percentage of correct predictions against all the others. However, it is not enough for binary classification problems where usually the classes are imbalanced. Precision was computed to see how precise the model is in filtering out cases and correctly identifying the actual positive ones, like identifying cases dealing with patients having diabetes. Recall, on the other hand, gave us the metrics that allowed us to assess the sensitivity of the model to real-world cases regarding diabetes by measuring how well it could detect all the positive cases. Finally, the F1 score, a balanced metric that takes into account both false positives and false negatives, was used for giving a more fine-grained view of how well this model performed for situations where both precision and recall are critical.

## **Data Augmentation and Regularization**

Other data augmentation strategies were used, including horizontal flipping, width and height shifts, and random rotations, to enhance model generalization capabilities during training. In fact, such augmentations enhanced the robustness of the model against changes in input data by artificially inflating the size of the training set. Regularization methods, including dropout, prevented overfitting by ensuring meaningful patterns of data were learnt by the model instead of the training set.

**2.5 Hyperparameter Tuning**

Hyperparameter tuning is essentially the process of choosing a hyperparameter set that is considered optimal for the model. These are pre-training parameters not learned from data but are extremely important to get the best performance out of any machine learning model. The most important hyperparameters in this research, including learning rate, batch size, and the number of epochs, have been tuned by trial and error. Most of the tests also used a learning rate of 0.001 because that is the factor that controls the rate of update steps for the model based on the gradient of the loss function. That number reached a good balance because it allowed continuing progress toward minimizing the loss function without updates becoming unpredictable. The batch size, which is the amount of samples the model processes before adjusting its weights, was set to 32. This magnitude made weight adjustments effective without being too onerous.

Early stopping was used to calculate the optimum number of epochs-that is, the number of iterations over the whole training dataset-in order to avoid overfitting. It stops when the model no longer improves after a certain number of training epochs. Thus, early stopping was used to terminate it and monitor the validation accuracy. In this way, one keeps the model away from overfitting because it was being trained on noise or unimportant patterns. It uses either a grid search or random search approach in systematically or arbitrarily testing several combinations of hyperparameter values for determining the model's best settings. Employing these hyperparameter tuning strategies balances the training time and predicted performance in ensuring the model would generalize successfully to new data.

# **Results**

The final results using tuned parameters provided substantive conclusions in logistic regression models. Here, the model fit was done using 5-fold cross-validation on 192 candidate parameter sets, totaling 960 fits. GridSearchCV best parameters were identified as alpha=0.01, initial learning rate (eta0) of 0.1, using the invscaling learning rate schedule, with a maximum of 1000 iterations, elasticnet regularization (penalty), random\_state set to 42.

Cross-validation provided a robust basis for model optimization, with an accuracy of 0.69 on the final performance in the test set, corresponding to a correct prediction of the outcome in 69% of the cases.

A screen shot of a computer

Description automatically generated

A blue squares with white text

Description automatically generated

1. **Discussion**

The logistic regression model yielded very poor success regarding classifying diabetes with an accuracy of 0.69 on the test set, even though features such as glucose level and BMI are so crucial in distinguishing between a diabetic patient and a non-diabetic. This model suffers under generalization to unseen data. One major limitation was the restricted set of features in the dataset. It contains important health parameters, but does not include other factors that might complete the model, like family medical history or some specific habits related to diet and exercise, which would contribute to an even more accurate prediction. Including these features will further increase the model performance.

Besides, it is highly imbalanced, as the non-diabetic cases outnumber the diabetic ones. This must have biasedly affected the learning process of the model, hence poor prediction for diabetic patients. The confusion matrix and classification report also showed evidence of strong bias in predicting non-diabetic cases which badly affects the overall performance. It could be resolved by using resampling techniques or re-tuning of the model parameters for better performance.

# **Conclusion**

The idea of this project is that logistic regression can predict the occurrence of diabetes effectively based on patient data. This model, as described above, worked well by keeping good accuracy along with features such as glucose, BMI, and age. Future effort may be made in increasing the number of features and then trying advanced models like random forests or neural networks.

# **References:**

 I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.

 F. Pedregosa, G. Varoquaux, A. Gramfort, et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.

 D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in Proceedings of the 3rd International Conference for Learning Representations (ICLR), 2015.

 H. Zhang, M. Cisse, Y. Dauphin, and D. Lopez-Paz, "mixup: Beyond Empirical Risk Minimization," in Proceedings of the 6th International Conference on Learning Representations (ICLR), 2018.

 Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-Based Learning Applied to Document Recognition," Proceedings of the IEEE, vol. 86, no. 11, pp. 2278-2324, Nov. 1998.

 P. J. Werbos, "Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences," Ph.D. dissertation, Harvard University, 1974.