DIABETIC RETINOPATHY MACHINE LEARNING MODEL

Coursework 1

Student Name: Furkan Baki

UB Number: 19001214

Contents

1. Introduction 2
1.1. Diabetic retinopathy 2
2. Background 2
3. Methodology and Data 2
3.1. Data 2
3.2. Logistic Regression 3
4. Analysis and Discussions 3
5. Conclusions and future work 3
6. Glossary 4
7. Bibliography 4

1. Introduction

1.1. Diabetic retinopathy

Diabetic retinopathy (DR) is a type of diabetes that affects the back of the eye (the retina). Its primary cause is high blood sugar levels. If left undiagnosed and untreated, it can cause blindness. The purpose of this report is to document the use of the Diabetic Retinopathy Debrecen data set, alongside a machine learning model, to predict whether a patient shows signs of diabetic retinopathy or not.

This report will go through the process taken in creating this machine learning model, with explanations for: the nature of the data; the feature engineering that was done on the data; the machine learning techniques that were used to model the data (Linear/Logistic regression, decision trees etc.); the evaluation of the solution amongst other things.

2. Background

The field of artificial intelligence has a wide variety of applications. One specific application is in the field of medicine for the prediction of diabetes. Machine learning algorithms can be fed data to produce models which can be used to predict whether a patient has diabetes or some other underlying medical conditions, based on the medical data of the patient. This can be done via automated retinal screening as one example, where "deep learning algorithms have been developed to automate the diagnosis of diabetic retinopathy". The diagnosis of a patient with diabetes can be done automatically by a trained machine learning model, with a "high sensitivity and specificity of 92.3% and 93.7% respectively", for automated screening of the retina (Ellahham, 2020).

3. Methodology and Data

3.1. Data

The dataset used will be the Diabetic Retinopathy Debrecen Data Set. This multivariate dataset contains 1150 instances and 20 attributes. It contains features extracted from the Messidor image set, which is a database that contains 1200 images of diabetic retinopathy retinal scans. The dataset has the following attributes:

No.	Attribute Information
0	The binary result of quality assessment. 0 = bad quality. 1 = sufficient quality.
1	The binary result of pre-screening. 1 = severe retinal abnormality. 0 = lack thereof.
2 - 7	The results of MA (microaneurysms[1]) detection. Each feature value stands for the number of Mas found at the confidence levels alpha = 0.5, …, …, 1 respectively.
8 - 15	Same as 2 – 7, but for exudate[2] detection. These values have been normalized.
16	The Euclidean distance of the center of the macula and the center of the optic disc. These values have been normalized.
17	The diameter of the optic disc.
18	The binary result of the AM/FM-based classification.
19	Class label. 1 = contains signs of DR. 0 = no signs of DR.

(Classification from
http://archive.ics.uci.edu/ml/datasets/Diabetic+Retinopathy+Debrecen+Data+Set)

AM/FM-based classification is the use of multiscale amplitude-modulation frequency-modulation (AM-FM) method to discriminate between normal and pathological retinal images (Agurto et al., 2010).

3.2. Logistic Regression

The AI system that was chosen to model the data was logistic regression. This is because the dependent variable is a binary classifier, and logistic regression can be used to predict whether a patient has diabetes (1) based on their screening, or not (0). Logistic regression estimates the probability of an event occurring, such as having diabetes or not, based on the given dataset. In the context of machine learning, logistic regression is a supervised machine learning model. It can also be considered a discriminative model, which means that it attempts to distinguish between classes. (IBM, 2023).

4. Analysis and Discussions

Before the data was fed to the logistic regression, it had to be pre-processed. The data itself had values of 0 in the columns that did not have binary values. Therefore, it could be assumed that these values of 0 were meant to be empty, and they were simply set to 0 as a placeholder. In order to achieve an accurate model, these rows were removed. The data was then scaled and split into training and testing data, which were to be used in cross-validation of our model, and then the model was trained. The evaluation of the significance and accuracy of the model was then done after this, using the testing data.

5. Conclusions and future work

A logistic regression model was created with an accuracy score of 0.73, or 73% which is a decent score, and demonstrates that the model is significant. However, this differs from the 92.3% that was shown in the literature review. This may have been due to either the quality of the data, or the method that was employed.

A logistic regression curve was also plotted for the MA detection levels and the exudate detection levels. From these graphs, there was a clear correlation between the MA and exudate levels of a patient and their chance of having diabetic retinopathy.

For future work I will need to do more work on feature engineering of the data. I was not too sure if I should have removed certain features or performed principal component analysis (PCA) as all the features seemed to contribute to the detection of DR. I plan to develop a deeper understanding into machine learning so that I can better employ different machine learning models.

6. Glossary

1. Microaneurysm – Tiny bulges in that eye that appear in the blood vessels of the retina. These may leak small amounts of blood and are very common in people with diabetes.
2. Exudates – Fluid that leaks out of blood vessels into nearby tissues (also known as pus).

7. Bibliography

NHS. (n.d.). *Diabetic Retinopathy*. NHS choices. Retrieved March 15, 2023, from https://www.nhs.uk/conditions/diabetic-retinopathy/

Diabetic Retinopathy Debrecen Data Set. UCI Machine Learning Repository: Diabetic retinopathy debrecen data set. (n.d.). Retrieved March 15, 2023, from http://archive.ics.uci.edu/ml/datasets/Diabetic+Retinopathy+Debrecen+Data+Set

Ellahham, S. (2020). Artificial Intelligence: The future for diabetes care. *The American Journal of Medicine*, *133*(8), 895–900. https://doi.org/10.1016/j.amjmed.2020.03.033

Agurto, C., Murray, V., Barriga, E., Murillo, S., Pattichis, M., Davis, H., Russell, S., Abramoff, M., & Soliz, P. (2010). Multiscale AM-FM methods for diabetic retinopathy lesion detection. *IEEE Transactions on Medical Imaging*, *29*(2), 502–512. https://doi.org/10.1109/tmi.2009.2037146

What is logistic regression? IBM. (n.d.). Retrieved March 19, 2023, from https://www.ibm.com/uk-en/topics/logistic-regression#:~:text=Resources-,What%20is%20logistic%20regression%3F,given%20dataset%20of%20independent%20variables.