ANALYSIS OF ALGORITHMS FOR DIABETES RISK SCORE

Subrata Sahoo
(18MCA1072)
School of Computing Science and Engineering
Vellore Institute of Technology
Chennai, India

Dr. Anusha. K
School of Computing Science and Engineering
Vellore Institute of Technology
Chennai, India

Abstract— **Diabetes is an illness that is commonly undetected, for which screening is advised from time to time. This widespread undiagnosed disease has a number of risk factors are related to it. The objective of this study is to evaluate the effectiveness of the chance of a score in a massive and representative population using available facts. Aim is to develop a model for the prediction of type diabetes risk on the basis of a logistic model and perceptron using neural network.**

Keywords—Diabetes, Perceptron, logistic Regression

I. INTRODUCTION

Diabetes is a massive burden in healthcare worldwide. Studies on life-style changes and drug intervention have convincingly proven that some measures can save you from diabetes. Early identification of populations that have a high chance for acquiring diabetes is therefore critical for prevention strategies and is necessary to permit proper efforts to be taken for the huge variety of individuals at high threat, at the same time as avoiding the burden of prevention and treatment for the even large quantity of individuals at low hazard, each for the individual and for society. The expert exercise advisory committee of the Diabetes affiliation recommends screening for all overweight or obese adults (frame mass index (BMI) ≥25) of all age who've at least risk components for diabetes which incorporate family history or hypertension. The international Diabetes Federation advise recommend using a reliable, easy, and sensible hazard scoring machine or the use of logistic regression formulas, to calculate the opportunity of questionnaire to pick out people at excessive risk of diabetes.

II. METHODOLOGY

Perceptron and Logistic regression is used to classify the Pima diabetes dataset. There are 8 attributes taken in following system with the Pima diabetes dataset of 768 values. The prediction of the models are applied to an accuracy matrix.

The identification of subjects at increased risk of future disease accurately is fundamental for every prevention program. Variables inside the prediction models included, have been assessed with a baseline standard questionnaire for disorder records and lifestyle variables.

A. *Confusion Matrix*

A confusion matrix helps in understanding of how a classifier performs. It represents the actual and predicted classification which the classification system performed.

Structure of a confusion matrix.

		Predicted	
		Negative	Positive
Actual	Negative	p	q
	Positive	r	s

Where, p = number of predictions where the prediction is negative and the instance is actually negative, q = number of predictions where the prediction is positive but the instance is negative, r = number of predictions where the prediction is negative but the instance is positive, s = number of predictions where the prediction is positive and instance is actually positive.

Accuracy of the classifier (AC) = all out number of right prediction.

$$AC=\frac{p+q}{p+q+r+s}$$

True positive rate (TP) = positive cases which were accurately recognized

$$TP=\frac{s}{r+s}$$

False positive rate (FP) = negative cases which were inaccurately recognized as positive

$$FP=\frac{q}{p+q}$$

False negative rate (FN) = positive cases which were inaccurately recognized as negative

$$FN=\frac{r}{r+s}$$

True negative rate (TN) = negative cases which were accurately recognized

$$TN=\frac{p}{p+q}$$

B. ***PIMA diabetes dataset***

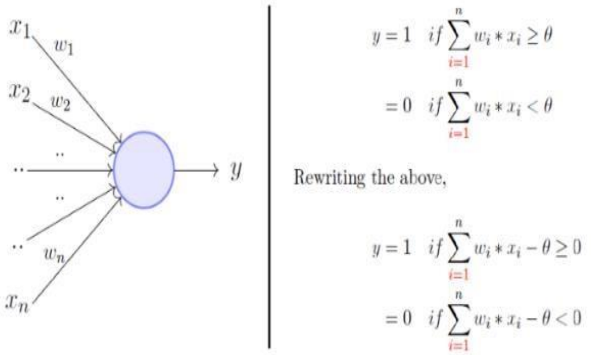
This dataset was created by the National Institute of Diabetes and Kidney Diseases to determine the prevalence of diabetes in the patient, based on the diagnostic characteristics of the dataset. The selection of these instances from a larger database went through a number of constraints. Particularly, majority of the patients are females who are at least 21 years old from the Pima Indian heritage.

The dataset consist of a number of medical predictor attributes and one class variable, Outcome. The dataset has 768 records. The attributes of the dataset are presented as follows.

Feature	Description
Pregnancies	Number of times the individual was pregnant
Glucose	In 2 hours in an oral glucose tolerance test the Plasma glucose concentration
BloodPressure	Diastolic blood pressure(mm Hg)
SkinThickness	Skin fold thickness of Triceps(mm)
Insulin	2-hour serum insulin (mu U/ml)
BMI	Body mass index (weight in kg/(height in m)^2)
DiabetesPedigreeFunction	Diabetes pedigree function
Age	Age in years

C. ***Classification Models***

Perceptron – The perceptron version is a greater trendy computational model than McCulloch-Pitts neuron. It takes an input, aggregates it (weighted sum) and returns 1 best if the aggregated sum is more than a few thresholds else returns zero. Rewriting the edge as proven above and making it a consistent enter with a variable weight.



x_1
w_1
x_2
w_2
x_n
w_n
y
$y = 1 \quad if \sum_{i=1}^{n} w_i * x_i \geq \theta$
$= 0 \quad if \sum_{i=1}^{n} w_i * x_i < \theta$
Rewriting the above,
$y = 1 \quad if \sum_{i=1}^{n} w_i * x_i - \theta \geq 0$
$= 0 \quad if \sum_{i=1}^{n} w_i * x_i - \theta < 0$

A single perceptron can simplest be used to enforce linearly separable functions. It takes each real and Boolean inputs and pals a hard and fast of weights to them, along with a bias by examine the weights, we get the feature.

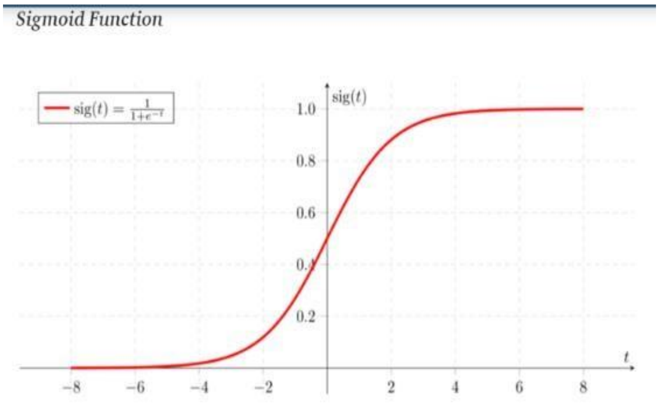
Logistic Regression – Logistic regression is the appropriate regression analysis to behavior when the established variable is dichotomous (binary). The logistic regression is a predictive evaluation, like any other regression analyses. Logistic regression is used to explain records and to give an explanation for the connection amongst one structured binary variable and one or extra nominal, ordinal or ratio-level independent variables.

Model

Output = 0 or 1 Hypothesis

=> Z = WX + B

hΘ(x) = sigmoid (Z)



III. EXPERIMENT RESULTS AND ANALYSIS

A. .***Experiment setup***

Python was used to implement the perceptron model and Logistic Regression model. The Pima diabetes dataset was split into training set and testing set. 80% of the Pima diabetes dataset was used to train the models.

B. ***Result analysis***

The rest of the 20% of the dataset was used to test the perceptron model and logistic regression model.

1) Confusion matrix

Confusion matrix for Perceptron model

		Predicted	
		Negative	Positive
Actual	Negative	102	5
	Positive	38	9

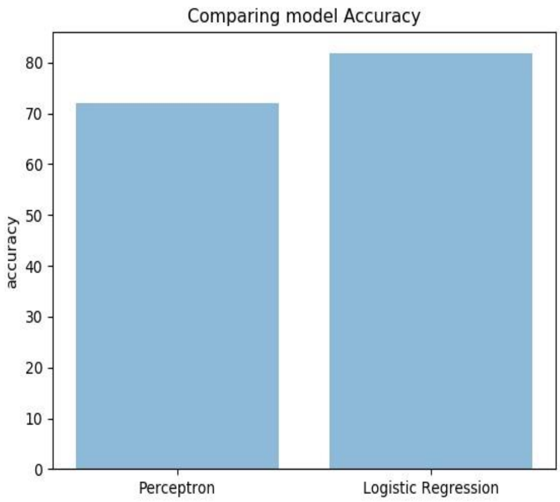
Confusion matrix for Logistic Regression model

		Predicted	
		Negative	Positive
Actual	Negative	98	9
	Positive	19	28

2) Accuracy

Comparison of the model

Model	Accuracy
Perceptron model	72.07
Logistic Regression model	81.81



The accuracy table shows that the Logistic Regression mode exhibited the higher (81.81) percent accuracy rate.

IV. CONCLUSION

Primary prediction models can identify people at excessive chance of growing diabetes in a time of 5 to 10 years. Models such as logistic regression and perceptron categorized instances, are barely better than simple the ones. Most models overestimated the actual hazard of diabetes. Current prediction models consequently perform to discover the ones at high threat, but cannot sufficiently quantify actual risk of future diabetes.