Data Normalization and Standardization: Impacting Classification Model Accuracy

Mani Butwall
Computer Science Department, ITM (SLS) Baroda University
Vadodara, Gujarat, India

ABSTRACT
In this paper, it was aimed to see the impact of the data normalization on the accuracy of classification model. In first part of this paper, the structure of dataset, features and basic statistical analysis of the data is represented. In this research, the study is done with the medical data set about the patients with the Diabetic disease. In second part of this paper, we present the process of data normalization and the impact of scaling data on the classification model performance. In this research, Deep Learning model is used for classification purpose. The main classification task was to classify whether the patient is diabetic or non-diabetic. Since the data set contains more numerical parameters of different scaling, the main aim of this paper was to investigate the impact of the data normalization (scaling) on the performance of the classification model. The purpose of the study is to show the difference in accuracy achieved by classification model with and without the use of scaling or normalization.

Keywords
Normalization, Classification, Diabetes Mellitus

1. INTRODUCTION
Data Mining and analytics has wide scope in the field of medical science nowadays. Machine learning methods for classification provide inexpensive means to perform diagnosis, prognosis, or detection of certain outcomes in health care research. Diabetes is disease in which the body does not properly produce insulin. It is one of the most common chronic disease, which can lead to serious long term complications. Type I diabetes, the disease is caused by the failure of pancreas, to produce sufficient insulin, and type II the body is resistant to the insulin it makes, which leads to an uncontrolled increase of blood glucose unless the patient uses insulin or drug. The blood glucose level that is elevated for a long period can result in metabolic complications such as kidney failure, blindness, and an increased chance of heart attacks. To prevent or postpone such complications strict control over the diabetic blood glucose level is needed [2].

The goal of this study is to propose various statistical normalization procedures to improve the classification accuracy. The experimental results showed that the performance of the diabetes classification model is enhanced when training of model is done with normalization. The result also shows the differences in the accuracy achieved with normalization and without normalization. The research is done with the sequential model of keras library. The Pima Indian Diabetes data set is taken for this study.

2. LITERATURE SURVEY
Statistics given by the Centers for Disease Control states that 26.9% of the population affected by diabetes are people whose age is greater than 65, 11.8% of all men aged 20 years or older are affected by diabetes and 10.8% of all women aged 20 years or older are affected by diabetes. The dataset used for analysis and modeling has 50784 records with 37 variables. After computational calculation they found 34% of the population with age<20 years and 33.9% of the population with 20<age<45 was not affected by diabetes. 26.8% of the population with age>45 years was not diabetic [7].According to another research work, used the Pima Indian diabetic database (PIDD) at the UCI Machine Learning Lab has been tested data mining algorithms to predict their accuracy in diabetic status from the 8 variables given. Out of 392 complete cases, guessing all are non-diabetic gives an accuracy of 65.1%. Rough sets as a data mining predictive tool applied rough sets to PIDD using ROSETTA software. The accuracy of predicting diabetic status on the PIDD was 82.6% on the initial random sample, which exceeds the previously used machine learning algorithms that ranged from 66-81% [8]. They proposed a system which predicts attributes such as age, sex, blood pressure and blood sugar and the chances of a diabetic patient getting a heart disease. They used Naïve Bayes Classifier and data set collected from diabetic research institute in Chennai which contain records of about 500 patients. The WEKA tool was used for Data mining with 10 fold cross validation. They found most of the diabetic patients with high cholesterol values are in the age group of 45 – 55, have a body weight in the range of 60 – 71, have BP value of 148 or 230, have a Fasting value in the range of 102 – 135, have a PP value in the range of 88 – 107, and have a A1C value in the range of 7.7 – 9.6 [9]. One more research concluded that the women suffering from diabetes can be tracked by clustering and attribute oriented induction techniques and dataset was collected from National Institute of Diabetes, Digestive and Kidney Diseases. The results were evaluated in five different clusters denoting concentrations of the various attributes and the percentage of women suffering from diabetes and they show that 23% of the women suffering from diabetes fall in cluster-0, 5% fall in cluster-1, 23% fall in cluster-2, 8% in cluster-3 and 25% in cluster-3 [10]. The data mining approaches in Diabetes diagnosis yields a result of almost 91-92%.

3. DATABASE EXPLANATION
Title: Pima Indians Diabetes Database

Number of Instances: 768

Number of Attributes: 8 plus class

For Each Attribute: (all numeric-valued)

1. Number of times pregnant

2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test.

3. Diastolic blood pressure (mm Hg)

4. Triceps skin fold thickness (mm)

5. 2-Hour serum insulin (mu U/ml)

6. Body mass index (weight in kg/(height in m)^2)

7. Diabetes pedigree function

8. Age (years)

9. Class variable (0 or 1)

Missing Attribute Values: Yes

Class Distribution: (Class value 1 is interpreted as "tested positive for diabetes")

3.1 Data Analysis

In our paper, we are dealing with the medical data about the patients with the Diabetic disease. This data set consists of biomedical data and is divided into two main categories. Figure 1 shows the distribution of these two main categories across the whole data set. The categories are Non Diabetic people and patients with the Diabetic disease.

Class Distribution: (Class value 1 is interpreted as "tested positive for diabetes")

Table 1 : Class Distribution

Class Value	Number of instances
0	500
1	268

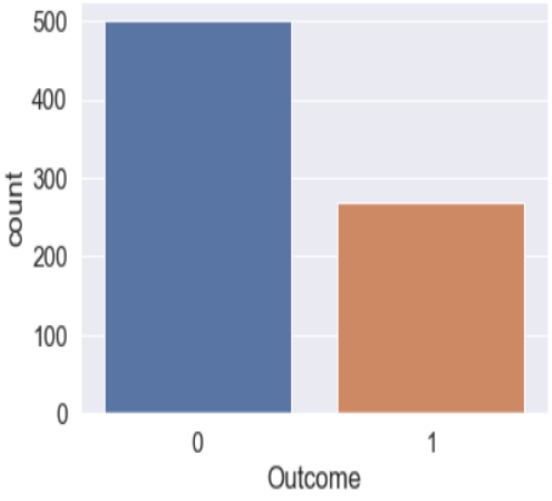

Figure 1: Class Distribution of Data

Before we start applying the data normalization and classification methods, it is needed to perform classical statistical analysis of the data. For each parameter, we computed statistical indicators like mean, standard deviation, minimum value, maximum value and quantiles. Figure 2 shows the computed values for each indicator.

	count	mean	std	min	25%	50%	75%	max
Pregnancies	768.0	3.845052	3.369578	0.000	1.00000	3.0000	6.00000	17.00
Glucose	768.0	120.894531	31.972618	0.000	99.00000	117.0000	140.25000	199.00
BloodPressure	768.0	69.105469	19.355807	0.000	62.00000	72.0000	80.00000	122.00
SkinThickness	768.0	20.536458	15.952218	0.000	0.00000	23.0000	32.00000	99.00
Insulin	768.0	79.799479	115.244002	0.000	0.00000	30.5000	127.25000	846.00
BMI	768.0	31.992578	7.884160	0.000	27.30000	32.0000	36.60000	67.10
iabetesPedigreeFunction	768.0	0.471876	0.331329	0.078	0.24375	0.3725	0.62625	2.42
Age	768.0	33.240885	11.760232	21.000	24.00000	29.0000	41.00000	81.00
Outcome	768.0	0.348958	0.476951	0.000	0.00000	0.0000	1.00000	1.00

Figure 2: Statistical indicators and character of the data set

4. NORMALIZATION

Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.

Let's say we have a dataset containing two variables: time traveled and distance covered. Time is measured in hours (e.g. 10, 15, 20 hours) and distance in miles (e.g. 200, 500, 1500 miles). Do you see the problem?

One obvious problem of course is that these two variables are measured in two different units — one in hours and the other in miles. The other problem — which is not obvious but if you take a closer look you'll find it — is the distribution of data, which is quite different in these two variables (both within and between variables).

The purpose of normalization is to transform data in a way that they are either dimensionless and/or have similar distributions. This process of normalization is known by other names such as standardization, feature scaling etc. Normalization is an essential step in data pre-processing in any machine learning application and model fitting.

Several methods are applied for normalization, three popular and widely used techniques are as follows:

- **Rescaling**: also known as "min-max normalization", it is the simplest of all methods and calculated as:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

- **Mean normalization:** This method uses the mean of the observations in the transformation process:

$$x' = \frac{x - average(x)}{\max(x) - \min(x)}$$

- **Z-score normalization:** Also known as standardization, this technique uses Z-score or "standard score". It is widely used in machine learning algorithms such as SVM and logistic regression:

$$z = \frac{x - \mu}{\sigma}$$

Here, z is the standard score, μ is the population mean and σ is the population standard deviation.

5. RESULT AND DISCUSSIONS

Data transformation and normalization improves the accuracy and efficiency of classification models. The purpose in this study is to see the effect in MSE and accuracy when applied to normalized data as compared to without normalized data.

Figure 3 shows the precision, recall, F1 score and support values without normalization and Figure 5 shows the precision, recall, F1 score and support values with normalization .

Figure 4 shows the MSE and Accuracy achieved without normalization and Figure 6 shows the MSE and Accuracy achieved on a normalized data.

	precision	recall	f1-score	support
0	0.78	0.70	0.74	159
1	0.46	0.56	0.50	72
accuracy			0.66	231
macro avg	0.62	0.63	0.62	231
weighted avg	0.68	0.66	0.67	231

```
[[112  47]
 [ 32  40]]
```

Figure 3: Precision, Recall, F1-Score and Support Values without Normalization

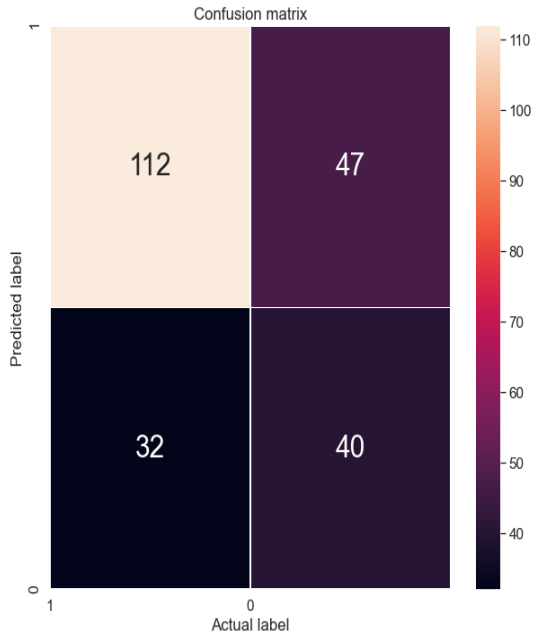

Figure 4: MSE and Accuracy without Normalization

	precision	recall	f1-score	support
0	1.00	1.00	1.00	159
1	1.00	1.00	1.00	72
accuracy			1.00	231
macro avg	1.00	1.00	1.00	231
weighted avg	1.00	1.00	1.00	231

```
[[159   0]
 [  0  72]]
```

Figure 5: Precision, Recall, F1-Score and Support Values with Normalization

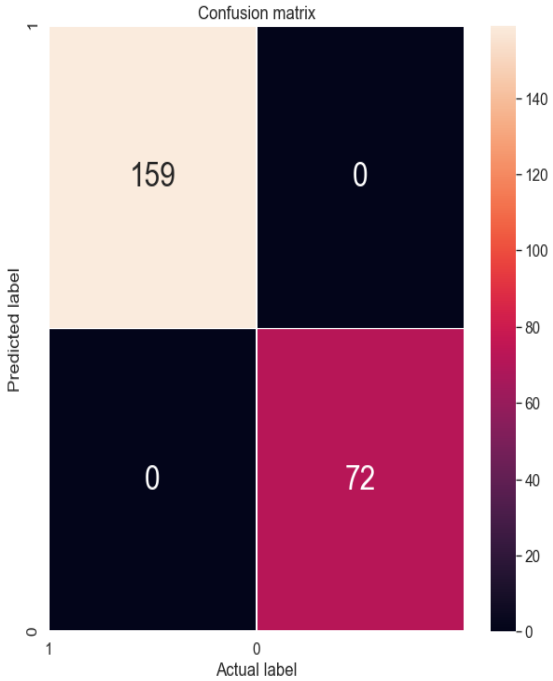

Figure 6: MSE and Accuracy with Normalization

Table 2: MSE and Accuracy

	Value Loss %	Value Accuracy %
With Normalization	**0.62**	**100**
Without Normalization	**72.30**	**65.80**

The accuracy value for the case where we used the raw data set was equal to 0.658. On the other hand, the accuracy value for the case where we used the normalized (scaled) data was equal to 0.100.

6. CONCLUSION

In this paper, we compared the accuracy of classification model in two cases. In the first case, raw data set was used with original values, and, in the second case, normalized data was used. The data after normalization was in same range of values. The main objective of this paper was to investigate the impact of the data normalization on the classification model accuracy.

However, this may be caused by a relatively small dataset or the character of the data. Since the results are not general, it can be useful to always investigate the accuracy parameter of the raw and normalized data.

7. REFERENCES

[1] Impact of Data Normalization on Classification Model Accuracy Dmitrii BORKIN1, Andrea NÉMETHOVÁ1, German MICHAĽČONOK1, Konstantin MAIOROV2 2019

[2] EngKhaledEskaf, Prof.Dr.Osama ,Badawi , Prof.Dr.TimRitchings,Predicting blood Glucose Levels in Diabetics using feature Extractionand Artificial Neural Networks.

[3] The Effect of the Normalization Method Used in Different Sample Sizes on the Success of Artificial Neural Network Model GökhanAksu 1*, CemOktayGüzeller 2, Mehmet TahaEser 2019

[4] 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", IEEE Symposium on Computer Applications and Medical Care, pp. 261-265, 1988.

[5] Statistical Normalization and Back Pro International Journal of Computer Theory and Engineering, Vol.3, No.1, February, 2011 1793-8201pagation for Classification T.Jayalakshmi, Dr.A.Santhakumaran

[6] Data Normalization to Accelerate Training for Linear Neural Net to Predict Tropical Cyclone Tracks Jian Jin,1 Ming Li,2 and Long Jin3

[7] Eurekalert, "Insufficient sleep may be linked to increased diabetes risk," July 11, 2010, <http://lifesciencelog.com/cluster53092620/.

[8] Bhatt K., Dalal P., Panwar A., "A Cluster Centres Initialization Method for Clustering Categorical Data Using Genetic Algorithm" International Journal of Digital Application & Contemporary research, 2013, Volume-2 Issue-1

[9] Huang, Zhexue, and Michael K. Ng. "A fuzzy k-modes algorithm for clustering categorical data." Fuzzy Systems, IEEE Transactions on 7.4 (1999): 446-452.

[10] Huang, Zhexue. "A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining."DMKD.1997.

[11] Importance of Input Data Normalization for the Application of Neural Networks to Complex Industrial ProblemsJ. Sola and J. SevillaDepartment of Electrical and Electronic Engineering Universidad Pública de Navarra. 31006 Pamplona,Spain