**EVALUATING THE PERFORMANCE OF FIVE MACHINE LEARNING ALGORITHMS USING ACCURACY AND ROC\_AUC\_SCORES.**

**Case study: The Pima Indians Database**

Mikel Kengni Ngueajio

Email: mikelkengni@gmail.com

Howard University, College of Engineering, Architecture and Computer sciences

Department of Computer Science, Washington, D.C 20059

**Abstract**:

Diabetes mellitus is one of the most serious health challenges in both developing and developed countries today. Diabetes is the seventh leading cause of death in the United States today, accounting for around $327 billion in total diagnosing costs in the US in 2017, $300 billion in direct medical cost and lost productivity each year. The CDC projects that one in three adults could have diabetes in 2050[source: <https://stateofobesity.org/diabetes/> ]. As a result, there is pressing need for developing classifiers and diabetes detectors at low cost and good performance, to effectively predict and properly treat diabetes.

The Pima Indian Diabetic Database which we picked from Kaggle (which is a platform for predictive modelling and analytics competitions and contains open datasets) is our case-study. The Pima Indian population have been under continuous study by the National Institute of Diabetes and Digestive and Kidney diseases(NIDDK) because of their high incidence rate for diabetes. In this paper, we predicted the target feature(outcomes) using 5 machine learning algorithms namely: - Logistic regression, Linear Support vector machine, Support vector machine with RBF kernel, Random forest classifier and Neural networks. To determine the performance of the different models, we used accuracy score and roc\_auc\_scores. The algorithm with an overall best accuracy and ROC\_AUC score is chosen and used for the final prediction.

**Section 1: Introduction**

Diabetes is a common health problem among people of different ages, genders and races. It is one of the fastest growing health problem in most countries and has no cure yet. Fortunately, Diabetes can be treated and controlled, and some people have even gone into remission from diabetes, by making little changes in their eating and exercising habits.

Insulin is a hormone made by the pancreas and it helps regulate the amount of glucose from blood getting into your cells, to be used for energy. Blood glucose is our main source of energy and it comes from the food we eat. An individual is said to be diabetic, when his blood sugar (also called blood glucose) level is too high as a result of your body not making enough (or any) Insulin or doesn’t use the insulin produces by your body effectively. The 2 most common types of Diabetes are Type 1 Diabetes and Type 2 Diabetes.

An individual with Type 1 Diabetes does not produce Insulin because his immune system attacks and destroys the cells in the pancreas responsible for making insulin. Type 1 diabetes is usually diagnosed in children and young adults and account for about 5 to 10 % of people with diabetes.

Type 2 diabetes also called Diabetes Mellitus is the most common type of diabetes and the major contributor to the high mortality rates due to diabetes. Unlike Type 1 Diabetes, Type 2 diabetic patients do not make or use insulin well and type 2 diabetes occur most often in middle-aged and older people. [ source: <https://www.niddk.nih.gov/health-information/diabetes/overview/what-is-diabetes> ]

As information systems in medical institutions get bigger, it becomes inefficient, difficult to manage and to extract relevant information from them. As a result, there is more need for automation in the health industry and Machine learning techniques have been considered to design automatic diabetes detection and diagnosis systems while providing better diagnosis accuracy, lower cost and human resources, and most importantly, low human errors.

In this study, we took a look at the performance of 5 machine learning algorithms namely: - Logistic Regression, Linear Support vector machines, Support vector machine with RBF kernel, Random Forest and Neural networks. First, we trained and validated the whole data on all 5 algorithms. Then we picked out the most important features using the random forest classifier feature importance parameter. We then trained and validated our new dataset, comprising these most important features together with the target feature(Outcome) on the 5 algorithms. We evaluated the performance of each of the machine learning algorithms using accuracy\_score and Roc\_auc\_score and the algorithm with the overall best score was used for our final prediction.

There have been many methods and algorithms used to approach the same problem. For example, in 1960, smith and all performed an evaluation using ADAP, to predict the onset of type 2 diabetes in the Pima Indian population. They used 576 randomly chosen data for training and 192 for testing. They reached an accuracy of 76% and they argued that “The neural network approach would provide strong results when the sample size is small, the form of underlying functional relationship is not known and the underlying functional relationships involve complex interactions and intercorrelations among a number of variables” [source : <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2245318/>]. On the other hand, Sahan and Al, used Attribute Weighted Artificial Immune System with 10 fold cross validation method and obtained an accuracy of 75.9% [ source: <https://link.springer.com/chapter/10.1007/11536444_35> ]. V. Anuja Kumari and Chitra used a 10 folds cross validation on the support vector machine with RBF kernel algorithm. They used 200 training samples and 260 testing samples got a 78% accuracy [ source: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.435.3161&rep=rep1&type=pdf> ].

This paper is organized as follows: Section 2 reviews some basic concepts on logistic regression, support vector machines, random forest classifier and neural networks, Section 3 describes the database and describes how the training, validation and performance evaluation for the 5 algorithms were undertaken. The experimental results are given in section 4. Finally, section 5 concludes the paper

**Section 2: Brief review of Proposed Algorithms**

1. **Logistic regression:**

Thisis a probabilistic and statistical classifier used to describe data and explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval independent variables. It helps predict the outcome of a categorical variable based on one or more predictor variable.

Logistic regression is defined as

http://www.statisticssolutions.com/wp-content/uploads/2010/01/log23.jpg

1. **Support Vector machines:**

SVMs are among the best “off-the-shelf” supervised learning algorithms( source: <http://cs229.stanford.edu/notes/cs229-notes3.pdf> ]. SVMs are a set of supervised learning methods used for classification, regression and outlier detection. They have the advantage that they are effective in high dimensional spaces, even in cases where the number of dimension are greater than the number of samples. The following implementations of the support vector classifiers were used in this paper:

1. *Linear:*

Linear support vector machine means a linear kernel was used in the support vector machine model.

1. *With RBF Kernel:*

The RBF kernel is the most popular choice of kernel types used in support vector machines mainly because of their localized and finite responses across the entire range of the real x-axis[source : <http://www.statsoft.com/Textbook/Support-Vector-Machines> ]. The output of the kernel is dependent on the Euclidean distance between xj and xi, one of which will be considered as the support vector and the other as the testing data point.

1. **Random Forest Classifier:**

A random forest is a meta estimator that fits a number of decision trees classifiers on various sub-samples of the dataset and makes a prediction by averaging the predictions of each component trees, to improve the predictive accuracy and control overfitting.

1. **Deep learning Neural Networks:**

The concept of Deep learning is a fast-growing technique, used in a range of diversified fields. In this paper, we use a single layered neural net with stochastic gradient descent as our optimizer, with a learning rate of 0.005. We train the neural network on 429 samples and validate on 108 samples, over 900 epochs (an epoch is defined as a single pass through all the training data).

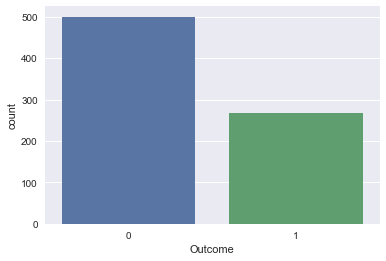
**Section 3: Predictive studies using all 5 algorithms**

1. **Experimental setup:**

The experiment was conducted on Jupyter Notebook. The datasets were downloaded from the Kaggle Competition interface, and stored as a .csv file. The dataset was directly read into Jupyter Notebook using Pandas. The performance of all the algorithms is evaluated using the accuracy\_score and roc\_auc\_score from SKLEARN. A ROC\_curve (true positive rates verses false positive rates) is also plotted for the different algorithms.

1. **Pima Diabetes dataset:**

The Pima diabetes dataset is originally from the National Institute Of Diabetes and Digestive and Kidney Diseases(NIDDK). It is a collection of 8 medical predictor variables and a target variable (Outcome), recorded from 768 female patients at least 21 years of age and of Pima Indian descent. The target variable(Outcome) takes variable 0- for non-diabetic patients which accounts for about 65% of patients and 1 for diabetic patients accounting for 35% of patients in the dataset. Below is a distribution of diabetic vs non-diabetic patients in the dataset.



The Pima Diabetes dataset comprises 8 numerical variables:

* Number of times pregnant (Pregnancies)
* Plasma glucose concentration a 2 hours in an oral glucose tolerance test (Glucose)
* Diastolic blood pressure in mm Hg (BloodPressure)
* Triceps skin fold thickness in mm (SkinThickness)
* 2-Hour serum insulin in mu U/ml (Insulin)
* Body mass index measured as weight in kg/(height in m)^2 (mass) (BMI)
* Diabetes pedigree function (DiabetesPedigreeFunction)
* Age in years (Age)

Technically the dataset has not missing values but has a lot of zeros which might have been substituted for missing values. For example, 11 patients had a BMI of 0, 227 patients had a skin thickness of 0, which does not make much sense.

So, we replace all the zero-valued columns except the pregnancies column with the mean value from each column.

The Table(1) below shows some statistical analyses of the PID dataset before and after imputing zeros with the mean (except for the pregnancies column).

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Before replacing zero values columns | | | | |  | After replacing the zero valued columns | | | |
| Attributes number | mean | Standard deviation | Min/max | | mean | Standard deviation | Min/max | |
| 1. | 3.8 | 3.4 | 0 | 17.0 | 3.8 | 3.4 | 0.0 | 17.0 |
| 2. | 120.9 | 32.0 | 0 | 199.0 | 121.7 | 30.4 | 44.0 | 199.0 |
| 3. | 69.1 | 19.4 | 0 | 122.0 | 72.3 | 12.1 | 24.0 | 122.0 |
| 4. | 20.5 | 16.0 | 0 | 99.0 | 26.6 | 9.6 | 7.0 | 99.0 |
| 5. | 79.8 | 115.2 | 0 | 846.0 | 118.7 | 93.1 | 14.0 | 846.0 |
| 6. | 32.0 | 7.9 | 0 | 67.1 | 32.5 | 6.9 | 18.0 | 67.0 |
| 7. | 0.5 | 0.3 | 0.08 | 2.42 | 0.47 | 0.3 | 0.07 | 2.4 |
| 8. | 33.2 | 11.8 | 21 | 81 | 33.2 | 11.8 | 21.0 | 81.0 |

**Table(1)**

In table (1) above, imputing the zero values with the mean had the effect of reducing the ranges between the attributes. But, the range is still high, so we used standard scaling to standardize (Standardizes the features by removing the mean and scaling to unit variance) the features even more. Hence reducing the range even further. This will boost our models’ performances.

1. **Training, Testing and Evaluations:**

In this part, we used 2 different datasets: The original pima dataset and a new dataset that was built from the most important features and the outcome feature.

Learning the parameters of a prediction function and testing it on the same data is a methodological mistake because, it might pollute the test data. Consequently, result in a model that give a perfect score for the sample provided, but fail to predict on yet-unseen data. This phenomenon is called Overfitting. One way to avoid it, is to hold out part of the data as test set, and use the training set instead, to do all the experiment. This concept is called cross validation.

1. ***Evaluating the models on the whole dataset:***
2. *Using Logistic regression, support vector machines and random forests:*

In order to evaluate the robustness of our models, we used a 10 folds cross validation on our training data set. This partitioned out training data set into 10 equally sized portions, and we use 9 randomly selected portions of the 10, to train the data and the remaining portion for testing. This is done 10 times until each portion serves as a test set at some point, then we average the model against each fold and get the final model score. We use the test data when we are sure what model performed better during cross validation.

1. ***Evaluating the models on the most important features and the target feature:***

In this part, we used the Random forest classifier to select the most important features.

From table(2), we see that the 4 most important features where Glucose, BMI, Age and DiabetesPedigreeFunction.

|  |  |
| --- | --- |
| **Features** | **Percentage importance** |
| Glucose | 25.531204 |
| BMI | 16.161817 |
| Age | 13.963501 |
| DiabetesPedigreeFunction | 12.415163 |
| BloodPressure | 8.422233 |
| Pregnancies | 8.20388 |
| insulin | 7.778867 |
| SkinThickness | 7.703334 |
|  |  |

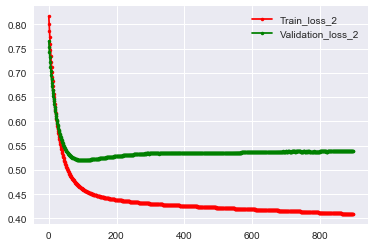
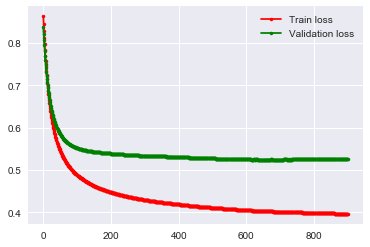
**Table (2)**

We created a new dataset comprising these 4 most important features and the target feature(Outcome) and then evaluated out models using a 10 folds cross validation (Exception neural networks).

1. **Using Neural network:**

In either part of this section, we used a single layered network of size 15, stochastic gradient descent as optimizer (with 0.005 as learning rate) and binary crossentropy for the loss function. We trained the data on 429 samples and validated on 108 samples over 900 epochs.

We got the following learning curves for both parts:



Learning curve before feature selection Learning Curve after feature selection

We also evaluated our model on another classification metric the roc\_auc score which is the Area Under the Roc curve. Roc curves were generated based on the predicted outcomes and they typically feature true positive rate(Sensitivity) on the y-axis and false positive rate(Specificity) on the x-axis. Sensitivity and specificity are statistical measures of how well our models discriminate between a case with positive and negative classes. The area under the ROC curve gives a measure of “how well are we separating the two classes” for this case, the bigger the roc\_auc score, the better.

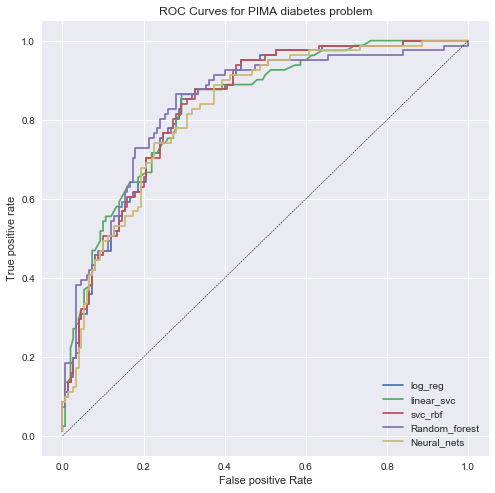
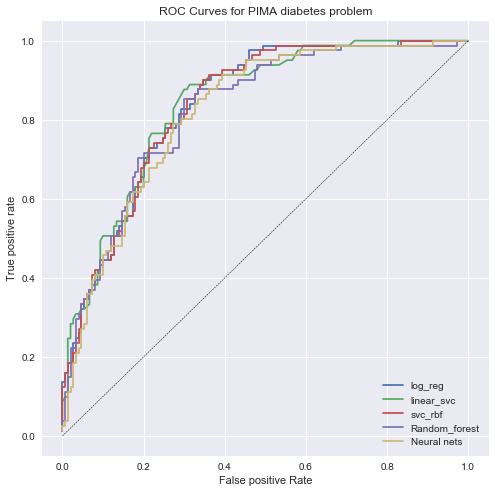
**Section 4: Result Analysis**

To analyze the performance of 4 of our models, we used a 10 folds cross validation as accuracy metric and a AUC measure. The experiment was conducted on 537 training data points and tested on 231 data point and the following results were obtained. Table 3 shows the models’ Accuracies before and after feature selection and table 4 shows the roc\_auc\_scores before and after feature selection.

|  |  |  |
| --- | --- | --- |
| Models | Model accuracies before features selection | Model accuracies after feature selections |
| Logistic regression | 77.638015 | 77.830189 |
| Linear support vector machines | 77.456324 | 77.830189 |
| Support vector machine with RBF kernel | 75.974843 | 76.897275 |
| Random forest classifier | 75.433263 | 76.530398 |
| Neural networks | 74.025974 | 74.458874 |

**Table (3)**

|  |  |  |
| --- | --- | --- |
| Models | Roc\_auc\_score before features selection | Roc\_auc\_score after feature selections |
| Logistic regression | 83.827160 | 83.432099 |
| Linear support vector machines | 83.802469 | 83.415638 |
| Support vector machine with RBF kernel | 82.543210 | 83.761317 |
| Random forest classifier | 84.493827 | 83.625514 |
| Neural networks | 82.090535 | 82.189300 |



Before feature selection After feature selection

Table 3 gives logistic regression as the model with the overall best accuracies of 77.8% after feature selection but from table 4, svm with rbf has the best roc\_auc\_score of 83.8%, just 0.3% greater than logistic regression. Overall, logistic regression is our best performing model.

**Conclusions:**

We used the Pima Indian Diabetes database form the Kaggle competition and the goal was to determine the outcome of the patients (Whether they are diabetic or not). These 5 machine learning algorithms namely: - Logistic Regression, support vector machine with RBF kernel, Linear support vector machine, random forest and Neural Networks were the min algorithms we chose, to predict the outcome. Four of them (mainly the Logistic Regression, support vector machine with RBF kernel, Linear support vector machine, random forest classifier algorithms) were evaluated using 10 folds cross validation while neural network was evaluated on accuracy. Their Roc\_auc\_scores were also calculated and logistic regression turn out to be the best performing model out of the 5 with an accuracy of 77.8 and a roc\_auc\_score of 83.4%.

However, it is important to consider the following limitations of the datasets, which might have produced a different outcome.

* The small size of the dataset may have limited the performance of some algorithms like Neural networks which work best with larger datasets
* The time the data was collected (around 1970s) may have not been as accurate as it would have been if the data were recently collected. Today, we have more sophisticated procedures for diagnosing diabetes like Urine test.

**References:**

1. Pima Indians Diabetes Database (Kaggle: website - <https://www.kaggle.com/uciml/pima-indians-diabetes-database> )
2. <https://stateofobesity.org/diabetes/>
3. <https://www.niddk.nih.gov/health-information/diabetes/overview/what-is-diabetes>
4. Cortes, C., Vapnik, V., “Support-vector networks”, Machine Learning, 20(2),pp. 273-297, 1995.
5. Jack W Smith, J.E. Everhart, W.C. Dickson, W.C. Knowler, R.S. Johannes, “Usind ADAP learning Algorithms to forecast the onset of Diabetes Mellitus” p. 245-318, 1988. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2245318/>
6. S.Sahan, K.Polat, H. Kodaz, and S. Gunes, “ The medical applications of attribute weighted artificial immune system (awais): Diagnosis of heart and diabetes diseases”, in ICARIS, 2005,p. 456-468 ( <https://link.springer.com/chapter/10.1007/11536444_35> ).
7. V.Anuja Kumari, R. Chitra, “Classification Of Diabetes Diseases Using Support Vector Machine”, Vol.3, Issue 2, International Journal Of Engineering Research and Application(IJERA), 2013, p.1797-1801 ”<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.435.3161&rep=rep1&type=pdf>
8. CS 229 Lecture notes from Stanford.edu – Andrew NG, Part V “Support Vector Machines” <http://cs229.stanford.edu/notes/cs229-notes3.pdf>
9. <http://www.statsoft.com/Textbook/Support-Vector-Machines>
10. Hung, et al. “Feature selection and classification model construction on type 2 diabetic patients‟ data”, Journal of Artificial Intelligence in Medicine, pp 251262, Elsevier, 2008.