Pima Indians Diabetes Dataset – Classification

Krishna Agarwal
School of Computing Sciences and Engineering, VIT Chennai, Tamil Nadu, India 600127
Email: krishna.agarwal2015@vit.ac.in

Abstract

The diabetes dataset is a binary classification problem where it needs to be analysed whether a patient is suffering from the disease or not on the basis of many available features in the dataset. Different methods and procedures of cleaning the data, feature extraction, feature engineering and algorithms to predict the onset of diabetes are used based for diagnostic measure on Pima Indians Diabetes Dataset [1].

Keywords

machine learning; Pima Indians Diabetes dataset; binary classification; features; feature extraction; feature engineering; support vector machine; MLP; neural netwroks; Decision tree; Linear regression heat map; pairplot; violin plot; feature importance.

1. Introduction

Pima Indian Diabetes dataset is to predict the onset of diabetes based on diagnostic measures. This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective is to predict whether a patient is suffering from diabetes or not based on diagnostic measurements. Classification algorithms include two main phases; in the first phase they try to find a model for the class attribute as a function of other variables of the datasets, and in the second phase, they apply previously designed model on the new and unseen datasets for determining the related class of each record Several constraints were placed on the selection of these instances from a larger dataset. Statistical classification is a problem studied in machine learning. It is a type of supervised learning, a method of machine learning where the categories are predefined, and is used to categorize new probabilistic observations into said categories. When there are only two categories the problem is known as statistical binary classification.[2]

The flow of the rest of the paper is as follows: In section 2, a little description about the dataset, its features and the outcome. In section 3, the methodologies that have been used in classification the algorithms implemented. In section 4, some other explanatory data analysis that have been done and results that have been found while implementing the various methodologies. In section 5, the conclusion that we establish from our experiments and the final result that we find. To conclude the paper is the citations from where reference has been made during the making of this paper.

2. Database - Pima Indians Diabetes Dataset

Pima Indian Diabetes dataset has 9 attributes in total. All the person in records are females and the number of pregnancies they have had has been recorded as the first attribute of the dataset. Second is the value of Plasma glucose concentration a 2 hours in an oral glucose tolerance test and then is the Diastolic blood pressure (mm Hg), fourth in line is the Triceps skin fold thickness (mm), then is the 2-Hour serum insulin (mu U/ml), sixth is Body mass index (weight in kg/ (height in m) ^2) and then seventh is the Diabetes pedigree function and the second last value is the that of the Age (years). The ninth column is that of the Class variable (0 or 1), 0 for no diabetes and 1 for the presence.

To start with we first take a description of the dataset. We infer not much from this except the facts like we have a data datset of 768 lines and the maximum values of the Age and Pregnancies. Nothing more is of much use for the prediction. We also calculated the number of datsets that were positive to the test of diabetes and those who were negetive and the value came out to be 268 and 500 respectively.

We decided to take the mean value of BMI and found that the average value of a person suffering from the disease has mean BMI value as 35.14 which means that they are not healthy and obese. It is also interesting to note that the mean BMI value for the people who are not suffering from the disease is 30 which is the threshold value of people becoming obese.[3]

The mean value of the second parameter Glucose (Plasma glucose concentration)[4] was done we found that those who suffered from the disease had mean value as 141.25 which indicates pre-diabetic state of hyperglycaemia that is associated with insulin resistance and increased risk of cardiovascular pathology.[5]

3. Methodology

A. A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements. Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal, but are also a popular tool in machine learning. [6]

Decision tree algorithm follows:

- The attribute/feature best for set is taken as root
- Distribute the set into different sets having same attribute values for particular value.

• Repeat the above steps till we get to the leaf nodes of the tree where no further division can take place.

B. In statistics, linear regression is a linear approach for modelling the relationship between a scalar dependent variable y and one or more explanatory variables (or independent variables) denoted X. The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regression. (This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable.) [7]

C. A multilayer perceptron (MLP) is a class of feedforward artificial neural network. An MLP consists of at least three layers of nodes. Except for the input nodes, each node is a neuron that uses a nonlinear activation function. MLP utilizes a supervised learning technique called backpropagation for training. Its multiple layers and non-linear activation distinguish MLP from a linear perceptron. It can distinguish data that is not linearly separable. [8]

D. In machine learning, support vector machines (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyse data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier (although methods such as Platt scaling exist to use SVM in a probabilistic classification setting). An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall. [9]

4. Experiments

To understand the data well we can do two things either we can just simply look at the data and check the values of each feature or we can do exploratory data analysis. If we choose the first option we would do the biggest sin ever in the history of a data scientist as we would become bias to the data and our mind inclines to the best know algorithm for the data but that would be good only for training. We may fail miserably in the testing data, hence we perform EDA.

So, next we start with plotting and the first plot that we perform is the pairplot.

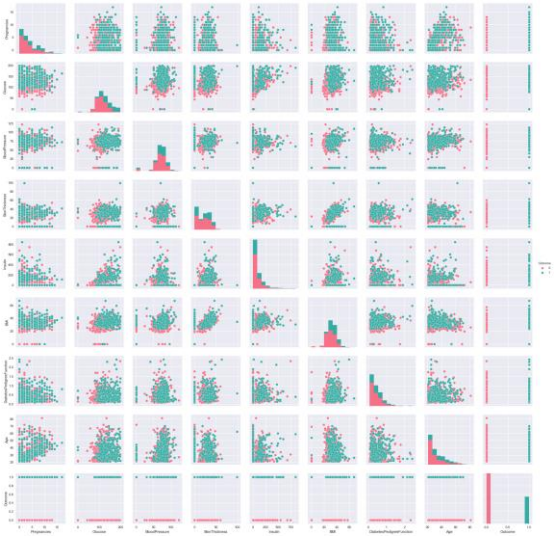

Figure 1 Pair-plot

One thing that we were able to deduce from this image was that all the parameters overlap for the Outcome value, i.e., no matter if you are suffering from the disease of not, you can have the same parameters.

Next in our list was the heat map plot which did give us some insight about the parameters and the relation it has with the other parameters and the Outcome as well.

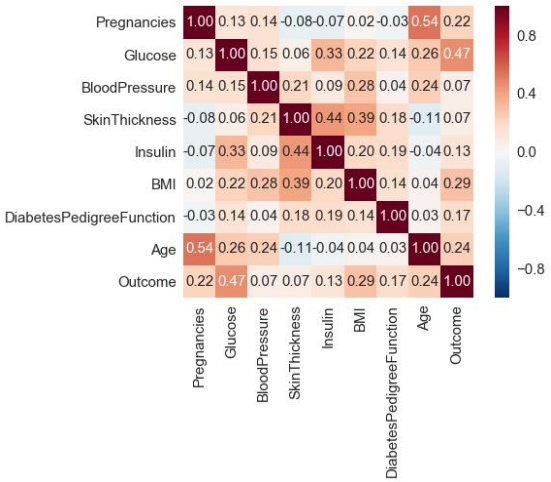

Figure 2 Heat Map

This heat map was of much use as we got to know that the following pairs had a positive correlation coefficient between them as compared to the other parameters:

1) Pregnancies and Age
2) Insulin and Skin thickness
3) BMI and Skin thickness
4) Insulin and Glucose

And with the Outcome value, Glucose and BMI values related the most.

This helped us to know that Glucose and BMI are the parameters we need to take special care of.

We further did some other EDA on some specific parameters viz., Glucose and BMI.

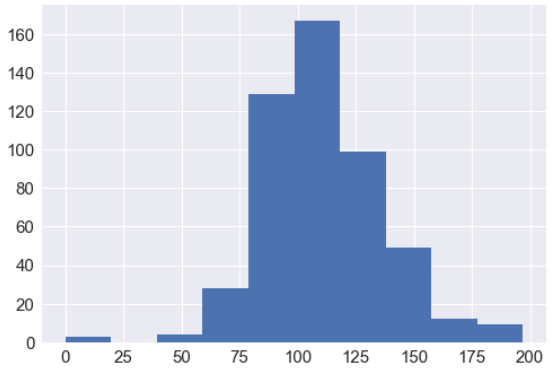

Figure 3: Histogram plot of Glucose value versus it's frequency in the dataset for negative Outcomes.

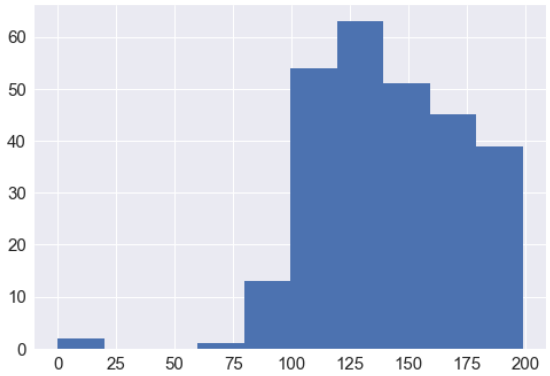

Figure 4: Histogram plot of Glucose value versus it's frequency in the dataset for positive Outcomes.

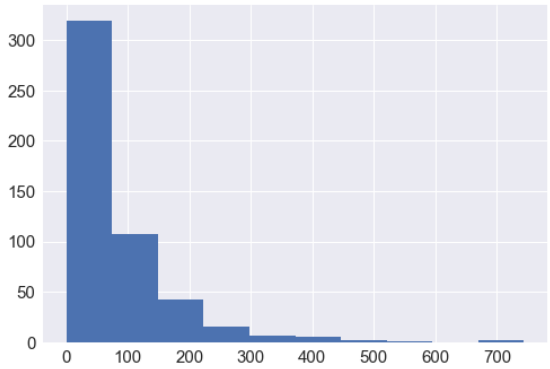

Figure 5: Histogram plot of Insulin value versus it's frequency in the dataset for negative Outcomes

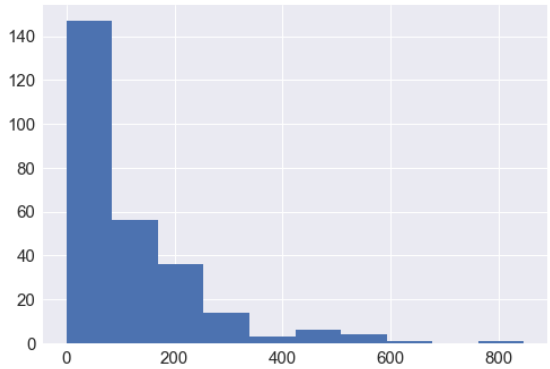

Figure 6: Histogram plot of Insulin value versus it's frequency in the dataset for positive Outcomes

We then wanted to see the distribution of the data points of all the parameters for the entire dataset therefore we plot the violin plots for positive and negative outcome separately.

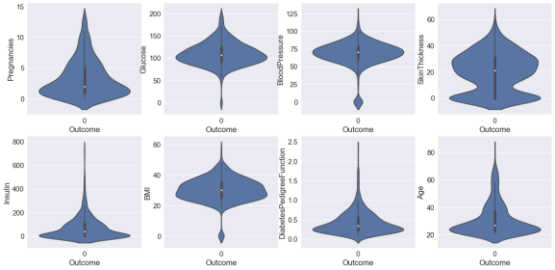

Figure 7: Violin plot for each parameter for Negative Outcome value

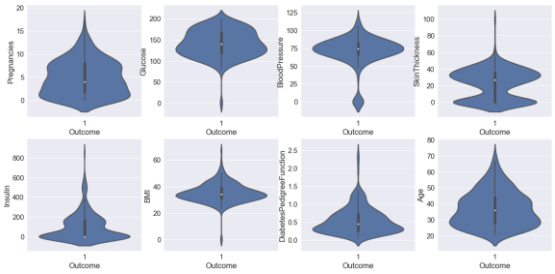

Figure 8: Violin plot for each parameter for Positive Outcome value

Now we knew for sure that Glucose, BMI and Insulin had the most effect on the Outcome value and so we will use these parameters while applying any algorithm.

Before going for any algorithm we divided our dataset into two parts training and testing in the 80:20 ratios.

To have a confirmation on the best features we found the feature importance using the Decision Tree. We also performed the prediction using the same algorithm which gave us an accuracy on training and testing data as 80.0 and 75.0 percentages respectively. We also plotted the feature importance plot and found Glucose and BMI to be the leading parameters.

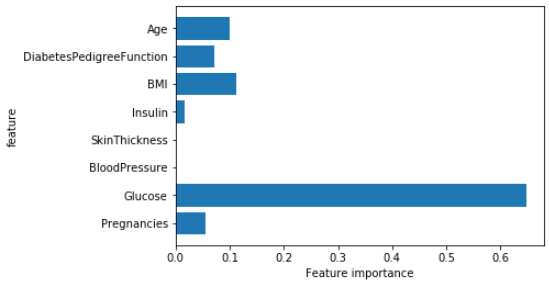

Figure 9: Feature Importance for each parameter

This was the decision tree that we got.

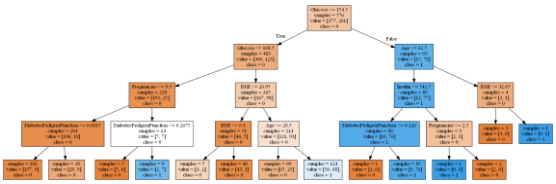

Figure 10: Decision Tree

Because we had found that instead of Insulin, Age was the other parameter that showed much importance so we did the plotting again with only three parameters in mind viz., Glucose, BMI and Age.

This was the feature importance plot:

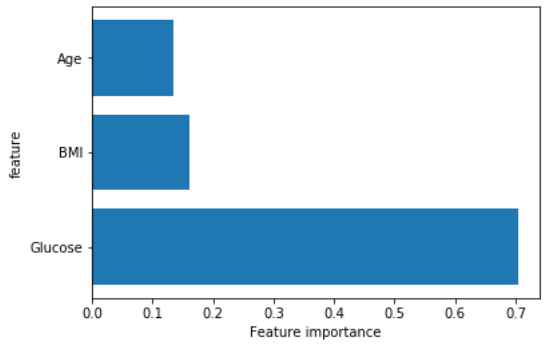

Figure 11 Feature importance plot for Age, BMI, and Glucose.

And this time the accuracy value for training and testing was 79.0 and 76.6 percentages respectively. This is better than the previous one as no matter the accuracy of the training has reduced but still for testing it came better.

We also used Linear Regression but not to predict the Outcome value because the accuracy was coming out very less moreover Linear regression is best for regression problems and not classification problems. The heat map showed the relation between each parameter and we saw that Age and Pregnancy had the best co relation. We used Linear Regression to predict the number of pregnancies that a woman would have had looking at her age. We trained our model and checked for different age values and the results were satisfactory.

To have a better prediction we went for MLP Neural Networks. We took the learning rate as 0.1 and ran our model 1000 times. This gave us a training accuracy of 71.66% and testing accuracy as 74.68%.

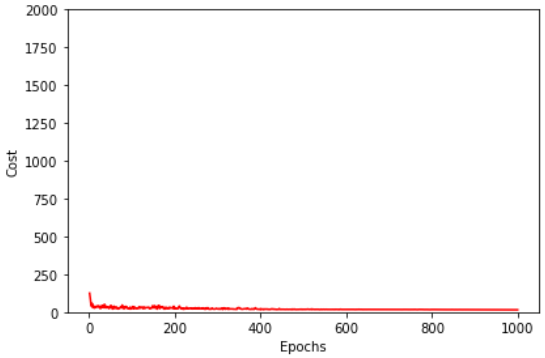

Figure 12: Epoch versus Cost graph

We can see in Figure 12 that the error approximated to zero. We can do even better than this by using General Regression Neural Networks. Kamer Kayaer et al[10] used it and got an accuracy of 80.21%. The highest true classification found on this dataset is of 81% using complex structured ARTMAP-IC network[11].

To have even better result we decided to go for pipelining and used the Support Vector Machine algorithm. The best score and parameter that we found were 0.75 using Linear Kernel. The testing accuracy achieved was of 80.5%.

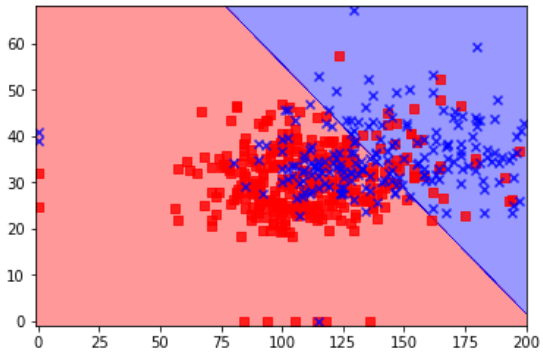

Figure 13: Support Vector Machine

5. Conclusion

We applied many algorithms and did a lot of feature manipulation and extraction. We got the best accuracy of 80.5% using SVM. A lot of information about the dataset was also extracted without using complex algorithms. We were also able to perform a lot of exploratory data analysis and came to many conclusions. Random Forest and Ensemble Learning can probably find a better result. Our result was also very close to the best result found and this shows that at the right parameters SVM can be a good and practical choice to classify a medical data.

6. References

1. https://archive.ics.uci.edu/ml/datasets/pima+indians+diabetes
2. https://en.wikipedia.org/wiki/Binary_classification
3. https://www.webmd.com/a-to-z-guides/body-mass-index-bmi-for-adults
4. https://en.wikipedia.org/wiki/Glucose_tolerance_test
5. https://en.wikipedia.org/wiki/Impaired_glucose_tolerance
6. https://en.wikipedia.org/wiki/Decision_tree
7. https://en.wikipedia.org/wiki/Linear_regression
8. https://en.wikipedia.org/wiki/Multilayer_perceptron
9. https://en.wikipedia.org/wiki/Support_vector_machine
10. https://www.google.co.in/url?sa=t&rct=j&q=&esrc=s&source=web&cd=4&ved=0ahUKEwjP9fSB6LXXAhUfSI8KHdKOBmYQFggg1MAM&url=http%3A%2F%2Fwww.yildiz.edu.tr%2F~tulay%2Fpublications%2FIcann-Iconip2003-2.pdf&usg=AOvVaw1sLNh9NoeKrIs4Xsg47QI4
11. http://techlab.bu.edu/resources/article_view/artmap-ic_and_medical_diagnosis_instance_counting_and_inconsistent_cases/