Analysing Feature Importances for Diabetes Prediction using Machine Learning

Debadri Dutta
School of Electronics
Kalinga Institute of Industrial Technology
Bhubaneswar, India
debadridtt@gmail.com

Debpriyo Paul
Computer Science Dept.
Institute of Engineering and Management
Kolkata, India
debpriyopaul96@gmail.com

Parthajeet Ghosh
School of Computer Science
Kalinga Institute of Industrial Technology
Bhubaneswar, India
piashghosh@gmail.com

*Abstract***—Diabetes is an uprising illness, particularly because of the kind of nourishment we are having these days and the conflicting eating regimen and schedule that we take after. Diabetes are fundamentally caused because of obesity or high glucose level, and so forth. So in this paper we will discover what are the critical elements for the reason for diabetes. Variable and feature choice have turned into the focal point of much research in regions of utilization for which datasets with tens or a huge number of factors are accessible.. Likewise we will center around the most essential features to predict whether a person will have chances to develop diabetes in the future.**

Keywords—Feature importance, data visualisation, machine learning, healthcare, predictive modelling, diabetes

I. INTRODUCTION

Diabetes mellitus has an immediate flag of high glucose, together with a few side effects including continuous pee, expanded thirst, expanded yearning and weight reduction. Patient of diabetes for the most part require consistent treatment, else, it will potentially prompt numerous perilous hazardous complications. The diabetes is determined to have the 2-hour post-stack plasma glucose being no less than 200mg/dL [1], and the need of recognizing diabetes convenient brings in different examinations about diabetes acknowledgment.

Numerous past research thinks about have been done about machine learning in diabetes recognizable proof. Research has been done centered around diabetes recognizable proof through SVM (Support Vector Machine) [2] and they acquired some rousing outcomes. Contrasting with the past work, we make a more comprehensive examination containing various regular systems used to diabetes ID, proposing to think about their execution and locate the best one among them.

Through this investigation, we look at a few normal and information preprocessors for every one of the classifiers we utilize, and locate the best preprocessor separately. At that point we think about using different classifiers after we adjust the parameters and different kernels of them to achieve their surmised most extreme precision.. Finally, we additionally investigate the pertinence of each element with the arrangement result, and this will change the informational index in future examinations.

Machine learning (ML), a sub-field of artificial intelligence,has advanced out of the need to instruct PCs how to automatically take in an answer for an issue. In designing this field is alluded to as example acknowledgement, relevantly named because the PC is separating patterns out of data and making a decision in view of the example identified. It is a rich field that is comprehensively and inalienably identified with flag handling most-notably through data driven learning systems [3,4] and it has rapidly gained its application in almost every field possible, starting from healthcare to product business, etc.

II. USE CASE AND ALGORITHMS USED

We experimented with three different algorithms to build our predictive model.The first algorithm that we used was Logistic Regression [5], next we used Support Vector Machines [6] and lastly we tried with Random Forest [7].

The algorithms were implemented on small Pima Indians of Arizona, diabetes dataset [8] which was retrieved by National Institute of Diabetes and Digestive and Kidney and had the diabetic tests for all the women of that region. Our main motive was to figure out the perfect predictive algorithm for diabetes classification of the women. We have also been able to figure out the most important features which play a major role in predicting Diabetes.

A. Logistic Regression

In the multiclass case, the preparation calculation utilizes the one-versus rest (OvR) plot if the 'multi_class' choice is set to 'ovr', and utilizes the cross-entropy misfortune if the 'multi_class' choice is set to 'multinomial'. (Presently the 'multinomial' alternative is bolstered just by the 'lbfgs', 'sag' and 'newton-cg' solvers.)

This class executes regularized calculated relapse utilizing the 'liblinear' library, 'newton-cg', 'list' and 'lbfgs' solvers. It can deal with both thick and scanty information. Utilize C-requested exhibits or CSR frameworks containing 64-bit glides for ideal execution; some other information arrangement will be changed over (and replicated).

The 'newton-cg', 'sag', and 'lbfgs' solvers bolster just L2 regularization with primal plan. The 'liblinear' solver underpins both L1 and L2 regularization, with a double definition just for the L2 regularization.

It uses a sigmoid function to predict the probability of a particular class, and if probability is >= 0.5, it is

assigned as class 1 and probability < 0.5 then it is assigned class 0. Below is the figure of a sigmoid function:

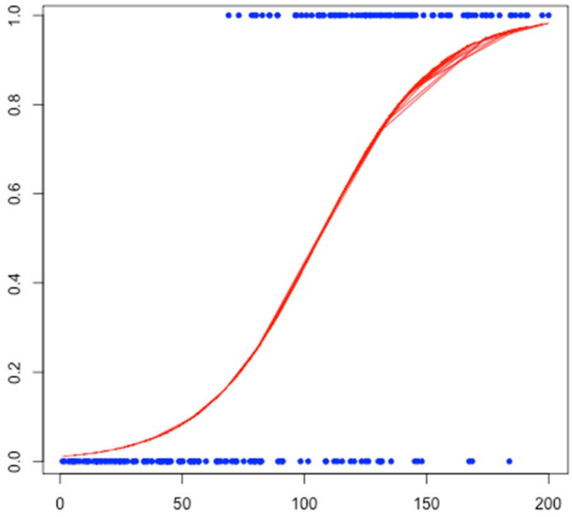


FIG. 1 SIGMOID FUNCTION

The formula for sigmoid function is given by:

$P = 1/1 + e^{-(a+bX)}$; 'P' being probability, 'a' and 'b' being the parameters of the model.

B. Support Vector Machines

A support vector machine builds a hyper-plane or set of hyperplanes in a high or limitless dimensional space, which can be utilized for classification, regression or different undertakings. Instinctively, a great separation is accomplished by the hyperplane that has the biggest separation to the closest training point of any class (supposed as functional margin), since all in all the bigger the margin the lower the generalisation error of the classifier.

The SVM has typically three kernels,

i) Linear Kernel

ii) Polynomial Kernel

iii) Radial Basis Function or Gaussian Kernel

Linear Kernels work perfectly on linearly separable data, however when data is linearly inseparable, RBF kernel is used which projects the data into an n-dimensional higher space and computes all the calculations there.

Below is the image of how a hyper-plane separates the data points of two different classes:

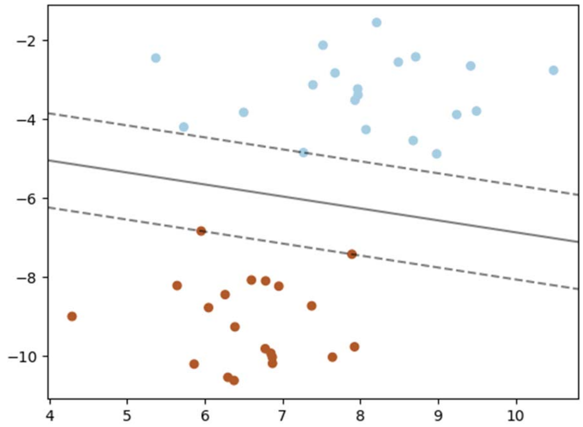


FIG. 2 SUPPORT VECTOR MACHINE HYPER-PLANE

The equation of a hyper-plane is given by:

$g(\underline{x}) = \underline{w}^t x + b$; where '$\underline{w}^t$'is the slope and 'b' is the y-intercept.
Support Vector Machine is also known as Maximal-Margin Classifier.

C. Random Forest Classifier

In random forests, each tree in the group is worked from an example drawn with substitution (i.e., a bootstrap sample) from the training set. Likewise, while splitting a node amid the construction of the tree, the split that is picked is not any more the best split among all highlights. Rather, the split that is picked is the best split among an random subset of the features. Because of this randomness, the bias of the forest marginally increases (as for the bias of a single non-random tree) in any case, because of averaging, its fluctuation likewise diminishes, generally more than making up for the expansion in predisposition, thus yielding a general better model.

The scikit-learn usage joins classifiers by averaging their probabilistic expectation, rather than giving every classifier a chance to vote in favor of a single class.

Below is an image of how the splitting looks like in a Random Forest:

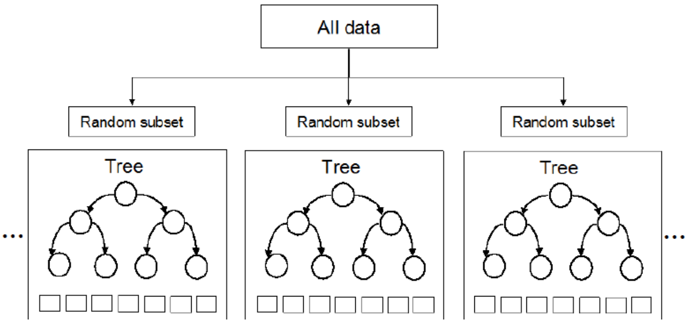


FIG. 3 RANDOM FOREST SPLITTING OF TREES

The random forest finds the best split using the Gin-Index Cost Function which is given by:

$Gini = \sum_{k=1}^{n} \quad (p_k * (1 - p_k)$; where k = each class and p = proportion of training instances.

D. Explanation of the Dataset and its Features

We made a model to predict whether the blood test detection for diabetes is going to turn out positive or negative. However ,the dataset was slightly imbalanced having around 262 class 0, i.e.class negative and 130 for class 1, i.e. positive. Below is an image of the class distribution.

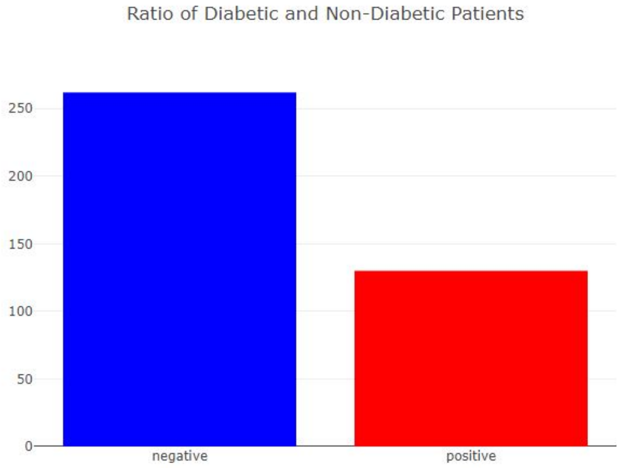


FIG. 4 DISTRIBUTION OF DIABETIC PATIENTS

First of all we divided the dataset into train and test in ratio of 67% train and 33% to test.

To fix the imbalancing we first used some upsampling techniques on the train dataset, to balance the two classes [9]. Also precision is not a good evaluation metric for imbalanced dataset, and we used f1-score and recall score [10] for evaluation which is discussed in the next section.

Following are the list of features that we had in our dataset:

i) **Pregnant:** It represents the number of times the woman got pregnant during her life.

ii) **Glucose:** It represents the plasma glucose concentration at 2 hours in an oral glucose tolerance test.

iii) **Diastolic:** The blood pressure is a very well-known way to measure the health of the heart of a person, there are too measure in fact, the diastolic and the systolic. In this data set, we have the diastolic which is in the fact the pressure in (mm/Hg) when the heart relaxed after the contraction.

iv) **Triceps:** It is a value used to estimate body fat (mm) which is measured on the right arm halfway between the olecranon process of the elbow and the acromial process of the scapula.

v) **Insulin:** It represents the rate of insulin 2 hours serum insulin (mu U/ml).

vi) **BMI:** It represents the Body Mass Index (weight in kg / (height in meters squared), and is an indicator of the health of a person.

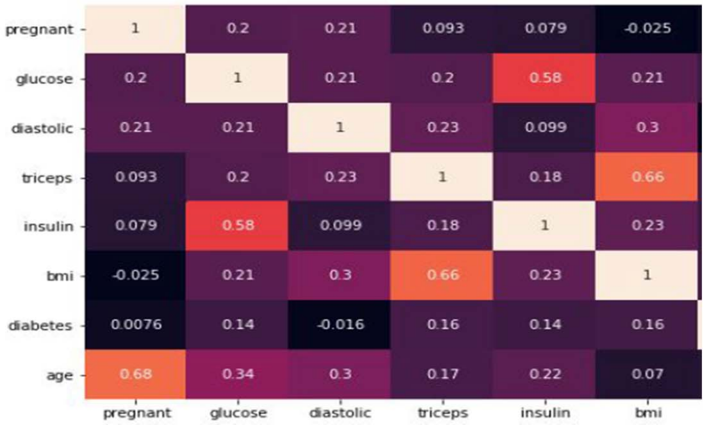
vii) **Diabetes:** It is an indicator of history of diabetes in the family.

viii) **Age:** It represents the age in years of the Pima's woman.

ix) **Test** (column to be predicted): It can take only 2 values ('negative' or 'positive') and represents if the patient shows signs of diabetes.

III. Calculations and Graphs

First of all we found the correlation between each feature columns, to check whether there is any highly correlated features, and as per our threshold of 0.7 there was none.. Below is the correlation plot:



The surprising fact is that, even on a small dataset, it was the bagging algorithm, Random Forest that surpassed all the other algorithms in the predictions. Following is a classification report, which signifies each of the algorithms' predictive accuracy, False +ve is also known as Type 1 error and False -ve is also known as Type 2 error. [11]

TABLE I. CONFUSION MATRIX

Name of Algorithm	True +ve	True -ve	False +ve	False -ve
Logistic Regression	73	64	18	18
Support Vector Machines	68	70	23	12
Random Forest	75	70	12	16

Following table is the f1-score and recall score for each class in case of Random Forest.

TABLE II. CLASSIFICATION REPORT (RANDOM FOREST)

FIG. 5 FEATURE CORRELATION PLOT

CLASS	F1-SCORE	RECALL
0 (NEGATIVE)	0.84	0.86
1 (POSITIVE)	0.83	0.81
AVG./TOTAL	0.84	0.84

The most important part for an algorithm to make predictions are the features which it is using for making the predictions, and some features play exceptionally important part for predicting. Below is the table, to how much importance has Random Forest put to each feature, following a graphical representation of the same.

TABLE III. FEATURE IMPORTANCE (RANDOM FOREST)

FEATURE NAME	IMPORTANCE
PREGNANT	0.07
GLUCOSE	0.21

DIASTOLIC	0.08
TRICEPS	0.09
INSULIN	0.19
BMI	0.12
DIABETES	0.11
AGE	0.13

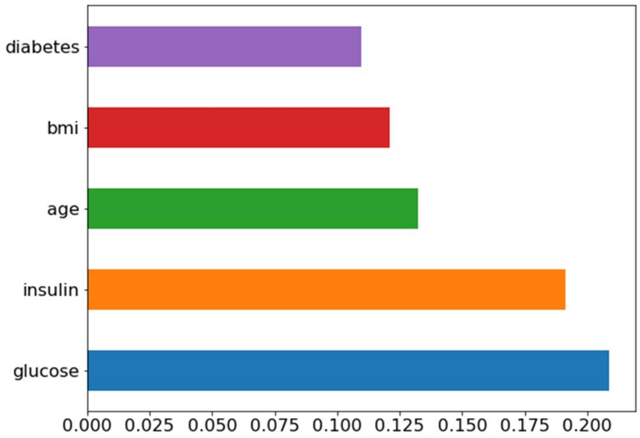


FIG. 5 FEATURE IMPORTANCE PLOT (RANDOM FOREST)

The sum of the importance of each feature will result to one. Above, only the features playing major role for diabetes have been plotted, where X-Axis represents the importance of each feature and Y-Axis the names of the features.

IV. CONCLUSION

So from the above calculations, we can conclude that Random Forest is the most ideal algorithm for predicting Diabetes, which gives an accuracy of around 84%. And if people want to prevent Diabetes, they should really keep their glucose level down and with increase in age they should follow a proper diet. Also people born in families having a diabetic history, they should really take care of themselves.

REFERENCES

[1] World Health Organization, “Report of a study group: Diabetes Mellitus,” World Health Organization Technical Report Series, Geneva, 727, 1985.

[2] Kemal Polat, Salih Gunes, and Ahmet Arslan, “A cascade learning system for classification of diabetes disease: Generalized Discriminant Analysis and Least Square Support Vector Machine,” Expert Systems with Applications, vol. 34. 1, January. 2008, pp. 482-487.

[3] Richard Duda, Peter Hart, and David Stork, PatternClassification, Wiley, New York, 2nd edition, 2001.

[4] Christopher M. Bishop, Pattern recognition and machine learning, Springer, New York, 1st edition, 2006

[5] Logistic Regression: http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

[6] Sci-Kit learn Support Vector Machines: http://scikit-learn.org/stable/modules/svm.html

[7] Random Forest: http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

[8] Pima Indians Diabetes Dataset: https://www.kaggle.com/mehdidag/pima-indians/home

[9] Sci-Kit learn Upsampling: http://scikit-learn.org/stable/modules/generated/sklearn.utils.resample.html

[10] f1-score and recall score: http://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_recall_fscore_support.html

[11] Type 1 and Type 2 error: https://www.cs.rpi.edu/~leen/misc-publications/SomeStatDefs.html