Optimized Machine Learning Ensemble for Diagnosis of Type II Diabetes

Pranshu Choubey
Information and Communication Technology
Manipal Institute of Technology, Manipal University
pranshuchoubey0@gmail.com

Sanidhya
Information and Communication Technology
Manipal Institute of Technology, Manipal University
sandisamp@gmail.com

Abstract—Ensemble methods in Machine Learning can open ways to more robust and enhanced methods for clinical studies and intuitive methods to diagnose some serious diseases. Ensemble techniques are used to achieve improved models having better predicting power and accuracy. This paper portrays the contrast between performance of ensemble techniques and simple classification techniques used to diagnose type II diabetes and it also provides a better and optimized machine learning ensemble model for detection of type II diabetes using electronic medical records of patients. The computation was strategically divided in two phases called Phase I and Phase II. Four machine learning classifiers have been used in Phase I: Naive Bayes, Logistic Regression, Random Forest and Simple CART on PIMA Indian datasets with 10-fold cross validation method. In Phase II two ensemble methods are used which are Simple Majority Voting (Voting) and Stacking Ensemble method. In phase I Linear Regression proved to be the best classification technique with AUC (Area under the curve) of ROC(Receiver operating characteristic) 0.832 and accuracy of 77.21% amongst all the four algorithms. In phase II the Simple Majority Voting gave the best result with AUC of ROC 0.836 and accuracy of 75.78% and on the parameter of AUC of ROC Simple Majority Voting optimized the diagnosis of Type II Diabetes in most significant way.

I. Introduction

The pervasiveness of type II diabetes and the affliction of disease caused by it have increased very hastily worldwide and ever growing. This has been driven by aging populations, meagre diet, and the concurrent epidemic of obesity [1]. Dietary factors increase the risk of developing type 2 diabetes by excess consumption of sugary drinks connected with amplified risks. The type of fats present in the diet play an important role, with saturated fats and trans fatty acids growing the risk, and polyunsaturated and mono-unsaturated fat diminishing the risk. In present scenario with existing clinical tools and methodologies, risk calculation of various diabetes generated diseases are not done efficiently. A risk calculation system with a decent accuracy for such diseases for a diabetic patient can be very useful for early prediction of those diseases. This paper proposes a model for type II diabetes risk diagnosis using accessible clinical and lab parameters on PIMA Indian Dataset. An accurate diagnosis or prediction algorithm could be very useful tool in medical advancement for this diagnosis. Normally, diagnosis of Type II diabetes patients is overdue 5-6 years after attack of the disease and by the time diagnosis results certify the patient of type II diabetes, they have established vascular complications, kidney problems and eyesight problems because of diabetes.

III. Methods And Techniques

The analysis was done in two phases named as Phase I and Phase II. In Phase I, a primary analysis was done in which the 10-fold cross validation of 9 features was used to train and test the data. In Phase 2 two ensemble techniques were introduced, Simple Majority Voting ensemble (Voting) and Stacked ensemble to provide a better classification model. The classification algorithms used in Phase I were Naive Bayes, Logistic Regression, Random Forest and Simple CART.

A. PHASE I

1) Naive Bayes: Naive Bayes is a Bayes Theorem based Machine Learning technique for constructing classifiers: models that allots class labels to problem instances, described as vectors of feature values, drawn from some fixed set. Naïve Bayes is not the only algorithm for training such classifiers, but a family of algorithms exists based on such common principle which states that, all Naive Bayes classifiers anticipate that the value of a precise feature is independent of the value of any additional feature, specified the class variable [4]. Bayes theorem delivers an approach of computing the subsequent probability, P(c|x) (posterior probability), from P(c) (class prior probability), P(x) (predictor prior probability), and P(x|c) (likelihood). Naive Bayes classifier assumes that the influence of the value of a predictor (x) on a certain class (c) is independent of the values of other predictors. This hypothesis is called class conditional independence.

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} \qquad (1)$$

2) **Logistic Regression**: The binary logistic model is used to approximate the probability of a binary response based on one or more predictor (or independent) variables (features) by approximating likelihoods with the help of a logistic function, which is the collective logistic distribution [5]. Thus, it uses analogous methods to treat the same set of problems as probabilistic regression, with the latter using a collective normal distribution curve instead. Equally, in the latent variable representation of these two methods, the first acquires a standard logistic distribution of errors and the second a standard normal distribution of incorrectness.

Logistic regression produces the coefficients (and its standard errors and significance levels) of a formula to anticipate a logit transformation of the likelihood of presence of the characteristic of interest[6].

$$logit(p) = b_0 + b_1X_1 + b_2X_2 + ... + b_kX_k \qquad (2)$$

where p is the likelihood of presence of the characteristic of interest. The logit transformation is defined as the logged
odds:

$$odds = \frac{probability\ of\ presence\ of\ characteristic}{probability\ of\ absence\ of\ characteristic} \qquad (3)$$

and

$$logit(p) = \ln\frac{p}{1-p} \qquad (4)$$

3) Random Forest: Random forests is a conception of compartmentalization, regression and other tasks, a general method of random decision forests that are an accumulation learning method that function by constructing a multitude of training time decision trees and produces the class - classification mode class or regression mean forecast of the distinct trees. Random decision forests are specific for decision trees' habit of over appropriate to their training set[7].
The random forest training algorithm applies the common procedure to tree learners of bootstrap aggregating, or bagging. Sanctioned a training set X = x1, ..., xn with result Y = y1, ..., yn, bagging repetitively (B times) picks a random sample with exchange of the training set and fits trees to these samples: For b = 1, ..., B: 1. Sample, with exchange, n training examples from X, Y; call these Xb, Yb.

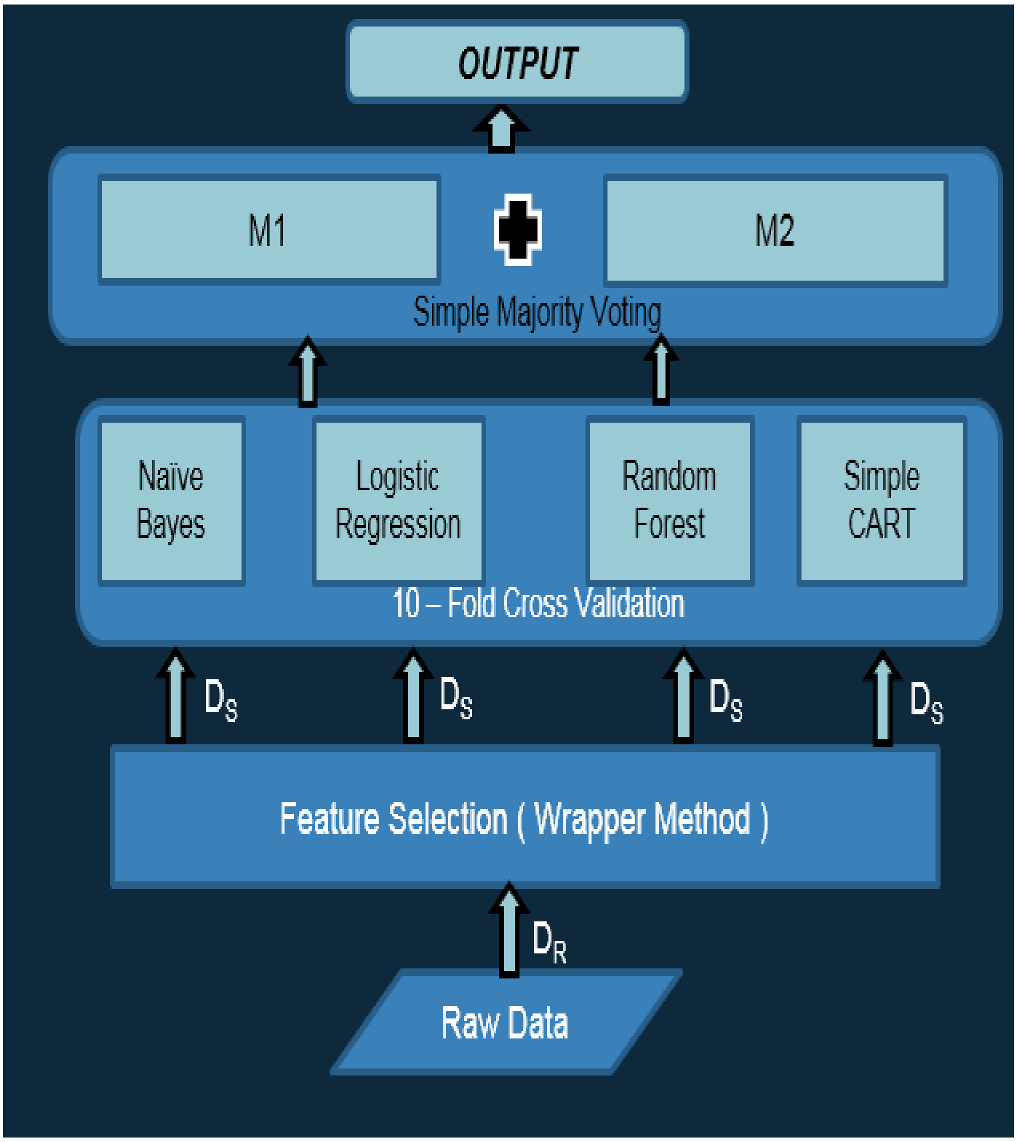
2. Train a decision or regression tree fb on Xb, Yb. After training, estimating for hidden samples x' can be made by averaging the estimates from all the individual regression trees on x':

$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B}\hat{f}_b\left(x'\right) \qquad (5)$$

4) Simple CART: Binary trees give an exciting and often enlightening way of looking at data in classification or regression problems [8]. Decision tree learning exercise decision tree as a projecting model which commute observations about an item to assessment about the item's target cost[9]. Objective variable in tree models can take a limited set of values and therefore tree is called classification trees. Tree leaves in these structures depict class labels and branches depict conjunctions of property that result to those class labels. Regression Trees are decision tree with uninterrupted valued target variable (typically real numbers) [10]. In decision analysis, decisions
and decision making are visually and explicitly represented using decision trees. A decision tree accounts data but not decisions in data mining; rather the resulting classification tree can be used as input for decision making. CART (classification and regression tree) algorithm uses Gini impurity that is an evaluation when an element is chosen randomly from the set of how often is an element incorrectly

labelled, if it were arbitrarily tagged per the distribution of labels in the subset. To reckon m items set Gini impurity, say, i {1; 2……m} and let f_i be the fraction of items labelled with value i in the set.

$$I_G(f) = \sum_{i=1}^{m} f_i(1 - f_i) \qquad (6)$$


OUTPUT
M1
M2
Simple Majority Voting
Naïve Bayes
Logistic Regression
Random Forest
Simple CART
10 – Fold Cross Validation
D_S
D_S
D_S
D_S
Feature Selection (Wrapper Method)
D_R
Raw Data