Proceedings of 9th IOE Graduate Conference
Peer Reviewed
ISSN: 2350-8914 (Online), 2350-8906 (Print)
Year: 2021 Month: March Volume: 9

An Ensemble Approach for the Diagnosis of Diabetes Mellitus Using Multiple Classifiers

Rachana Kunwar [a], Arun Kumar Timalsina [b]

[a, b] *Department of Electronics and Computer Engineering, Pulchowk Campus, IOE, TU, Nepal*
Corresponding Email: [a] 074mscsk009.rachana@pcampus.edu.np, [b] t.arun@ioe.edu.np

Abstract
According to the World Health Organization (WHO) 347 million people in the world are currently suffering from diabetes. WHO has also reported that diabetes will be the 7th prime cause of death in 2030 [1]. According to WHO, there is no exact data of patients with diabetes in Nepal. But, the estimated prevalence of T2DM in Nepal in 2015 was 8.4% in 2016 it was 9.1% and in 2017, it was found to be 11.7% [2]. This research has focused on developing an Ensemble model based on different base classification for diabetes mellitus diagnosis. Namely, Logistic Regression, Support Vector Machine, Naive Bayes, Decision Tree. The proposed ensemble method algorithm assembles the base classifiers with the probability of each individual classifier to attain the final result by computing the statistical mode for its output. Here, each single classifier gives the result for accuracy of 77.12%, 78.35%, 76.19%, 75.76% , 79.59% and 81.16% for Logistics regression, Support Vector Machine, Naive Bayes, Decision Tree, and the Ensemble method using vote without and with weight-age of 2 for Logistics regression and Support Vector Machine and 1 for NB and Decision Tree. The experimental result keeps up the idea that hybrid approaches are more implicit than the individual techniques of using classifiers separately.

Keywords
Support Vector Machine, Logistic Regression, Naive Bayes, Decision Tree, Voting Classifier, Ensemble Model, Weight-age, Cross Validation

1. Background

1.1 Diabetes Mellitus

Diabetes mellitus is an appalling disease due to more sugar in the circulatory system. The disease occurs if the pancreas does not generate proper amounts of insulin or when the body cannot essentially respond to the insulin that is produced. According to the World Health Organization (WHO) 347 million people in the world are currently suffering from diabetes.According to the research, more than 80% of diabetes deaths occur in low and middle income countries. WHO has also reported that diabetes will be the 7th prime cause of death in 2030 [**?**]. According to WHO, there is no exact data of patients with diabetes in Nepal. But, the estimated prevalence of T2DM in Nepal in 2015 was 8.4% in 2016 it was 9.1% and in 2017, it was found to be 11.7% [2].

If diabetes is ignored at an early stage, it may forefront to severe complications ending with death. Repetitive symptoms include increased thirst, frequent urination, and weight fluctuation. Moreover symptoms show up slowly, which incorporates loss of vision , diabetic neuropathy, Liver problems, Heart problems, etc[3]. Data mining in health sector is helpful for diagnosis with the help of different machine learning algorithms for finding the hidden pattern that enhances the accuracy rate and further prediction. [4]

Data mining is the analytical method for knowledge extraction from large databases. The tasks include data clustering, data association and data classification. The PID dataset has two outcomes, i.e either a diabetic patient or non diabetic patient . It's a binary classification problem where multiple individual classifiers are used at first and then combined with voting classifiers for results.

Ensemble methods is a new machine learning tool that combines different base models often called "weak learners" to give rise to one better assemble model.

The main theorem is that a combination of weak models can result in more accurate and/or powerful models.

The paper is divided into five chapters. Chapter 1, describes a brief background introduction to diabetes mellitus and the research objectives along with the problem statement. In Chapter 2, the relevant literature during this research studies considered and referred to in this research work has been included. Chapter 3, includes the methodology applied to this research work. It explains a method that is implemented for diabetes diagnosis. Chapter 4 includes the experimental results, evaluation and discussion on the proposed method. Chapter 5 includes the conclusion of the thesis and recommendations for future research in the area.

1.2 Research Background

Data Mining is applied on primarily two general approaches for the diagnosis of diabetes. A specific classification algorithm is the first way to assume the vulnerability on the patient's diabetes data. Hybrid algorithms are used as a second method.
The hybrid algorithm was implemented by Seyed and Razieh [3] to exemplify the ensemble algorithm with voting classifiers along with weight k– nearest neighbour, simple, decision trees and logistic regressions using Pima datasets. The system's accuracy is greatly increased by the hybrid approach.

Decision tree, naive Bayes, k-nearest neighbour & Pima Indian diabetes datasets tested by SVM were the different data mining algorithms demonstrated by Thirumal and Nagarajan [5].

An expert system was presented by Lee and Wang [6] that gave a semantic description of diabetes for diabetes diagnosis support application.

To lower the specific weakness of these algorithms i.e. SVM and Naive Bayes, a joint implementation of these two algorithms was proposed by Tafa et al [7]. Combining these algs, the accuracy of the method rose up to 87.6% which is a refinement. The negative false answers were decreased by the joint implementation which is an important achievement in medical diagnosis.

An SVM was used by Han et al. [8] in which SVM along with an ensemble learning module that treats the "blackbox" of the SVM decisions into logical rules to look for diabetic. The system which is hybrid was also found to be efficient and provided a tool for diabetes diagnosis according to the study.

2. Methodology

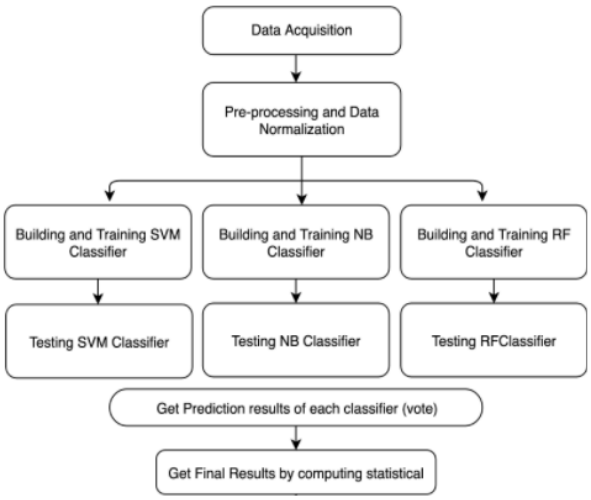

Figure 1: Proposed ensemble method algorithm

2.1 Data Acquisition

The Diabetes Dataset is accessible from UCI Machine Learning Repository for the download. The dataset with data from 768 women have 8 features instances and 1 outcome with binary classification:
Attribute Information

1. Number of times pregnant = Pregnancies
2. Plasma glucose concentration = Glucose
3. Diastolic blood pressure = BloodPressure
4. Triceps skinfold thickness = SkinThickness
5. Two-hour serum insulin = Insulin
6. Body mass index = BMI
7. Diabetes pedigree function = DiabetesPedigreeFunction
8. Age in years = Age
9. Class variable (0 or 1) = Outcome

Here, Outcome, is a binary classified Class with value 1 is defined as diabetic person and class value 0 is defined as "Non- Diabetic" patient.

2.2 Data Pre-Processing

2.2.1 Checking Data Types, Null values and zeros

The dataset consists of 768 records in total. All the values in the dataset are numeric. No Null values were

found in the datasets.

2.2.2 Data Cleaning

Some records are found with value "0" which is not possible for BMI, Insulin, Blood pressure, etc. except Pregnancies. So, while cleaning the dataset, we replace the 0 values in the specific column by calculating the average median value of the particular feature column. The median is the middle value from the ordered data set.

$$\left(\frac{n+1}{2}\right)^{th} \text{ value}$$

where n is the number of values in a data table. Here n=768, So, median value will be 384.5 which is not the exact cell. Thus, the median value can be calculated by averaging the values from 384th and 385th instances.

$$\text{Average} = \frac{\text{value below median} + \text{value above median}}{2}$$

2.2.3 Balancing the Data set

In total of 768, there are 500 counts for class 0, non-diabetes and 268 for class 1, diabetes.

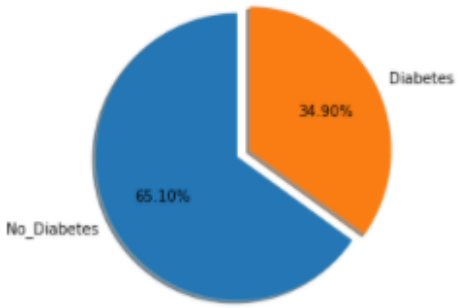

Figure 2: Outcome distribution of Diabetes Mellitus

2.2.4 Data Normalisation

Here, N Normalisation is done by using Maximum-Minimum scaling, the data is scaled to a fixed range [0, 1]. This will end up with smaller standard deviations. For all feature value X, new Xnorm can be calculated as:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

2.2.5 Correlation for all the features with Outcome.

As none of the features has a strong correlation ,i.e. Equals to 1 with Outcome, and are positively correlated with "Outcome" we take all the features in considerations.

2.2.6 Data Selection

For that the selection of data will be done in following manner:
1. Training set: 80%
2. Test set: 20%

2.3 Model Building

For model building, different classifications techniques namely Logistic regression, Support Vector Machine, Naive Bayes, Decision Tree were used to train the model . Grid search is used to hyper parameter tuning that will methodically build and evaluate a model for each combination of algorithm parameters specified in a grid.

2.3.1 Logistic Regression

The "S"curve from the logistic function indicates whether the patients have diabetes or not etc. It predicts the probability of the outcome that can only have two values that is 0 or 1. When output is 1 it means the value is greater than the threshold, else the output is 0. Since the logistic regression uses the standard logistic function interpreted as a probability which takes any real input x, (x $\in$ R), whereas the output always lies between 0 and 1.

Consider Y as a linear function of a single explanatory variable x. So, t equals:

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \ldots + b_nx_n$$

Here n = 8 feature samples.

And the logistic function is :

$$F(x) = y(1-y);\ 0 \ \text{ for } \ y = 0,$$

and infinity for y=1

For large range, then the logarithm of the equation become:

$$\log y(1-y) = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \ldots + b_8x_8$$

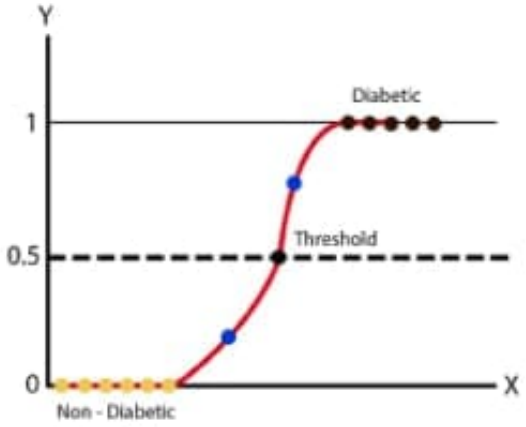

Figure 3: Logistic Regression

2.3.2 Support Vector Machine

To identify the right hyperplane.The equation to optimize maximum marginal hyperplane (MMH). distance is given by:

$$W^T x + b = -1$$

$$W^T x + b = 1$$

Maximizing the distances between neighbour data points. If the distance value is equal to

$$\frac{2}{||W||}$$

then it needs to be optimized again.

$$y_i * (W^T x_i + b) >= 1$$

, for all i $\in$ 1, $\not\subset\not\subset$,N. indicates that svm classifier to classify the class

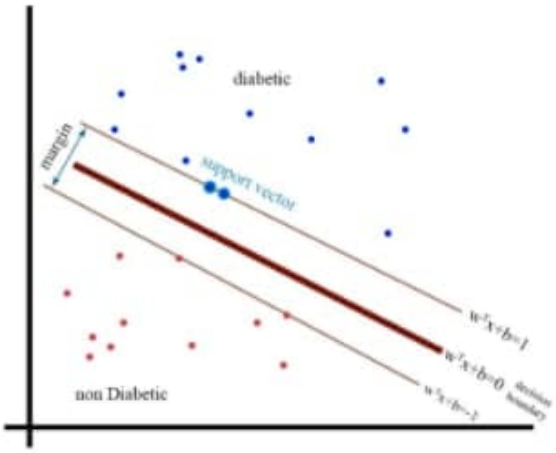

Figure 4: Support Vector Machine

2.3.3 Naive Bayes

Using Bayes theorem, posterior probability P(X| A) could be calculated from P(A| X), P(X) , and P(A).

$$P(X|A) = \frac{P(A|X) * P(X)}{P(A)}$$

Where,
P(X|A) - Posterior probability of target/object class.
P(A|X) - Predictor class probability
P(X) - True probability of Outcome - X
P(A) - Predictor prior probability

2.3.4 Decision Tree

In tree structures, branches describe conjunctions of features and leaves signify class labels. In this research, 8 features act as decision nodes in the tree with 2 leaf nodes classified with class labelled 0 and 1. Find the best attribute in the dataset using Gini criterion Divide the S into subsets that contain possible values for the best attributes.

2.3.5 Ensemble method

The proposed ensemble method algorithm applies Predicted value as a vote by each of the classifiers to attain the final result. This voting mechanism considers each estimation of the classifiers as an input to the ensemble system and then computes the statistical mode for its output to get the majority vote. Here four classifiers SVM, NB, LR and DT to predict the Final Prediction, Pf. Here two way of Voting classifiers:

2.3.5.1 Voting without Weights: Here all the classifiers results are treated equally. The Votes are calculated only in accordance to their mode value. This technique is also known as hard voting.

2.3.5.2 Voting with Weights: This is also known as soft voting. Here weightage of '2' to LR and SVM classifiers. Since these have better accuracy than the NB and DT which are weighted with value '1'. This technique of giving the weight helps in increasing the chances of not missing the cases that are diabetic is real. General formula for Voting

$$\hat{Y} = argmax \frac{1}{Nclassifiers} \sum (P1, P2, P3, \ldots, Pn)$$

Here N=4.

3. Models Performance Validation

3.1 Cross Validation Technique

This process repeats 5 times for the testing and training data sets and, finally, the error rates for 5 sets are averaged to yield an overall error rate and also estimated the predictive accuracy of the model trained with all the data for 5 times.

3.2 Confusions Matrix

This is used for the validation and performance evaluation of the model. The precision, recall and F1-score has been calculated. This gives a matrix as output and describes the complete performance of the model. This is used for the validation and performance evaluation of the model. The precision, recall and F1- score has been calculated. This gives a matrix as output and describes the complete performance of the model. Moreover, the diagonal cells indicate the percentage of predicted classes correctness while the off-diagonal cells represent the classifier mistakes.

4. Results and Discussion

This system has been experimented in google colab using Python language with UCI Machine Learning Repository diabetes dataset(PID). During this experiment we performed some data analysis tasks for data visualisation.

4.1 Dataset Visualisation

With a graphical visualisation of the data we have a better understanding of the various features values distribution. Some records have 0 values for some of the features, it's not possible to have 0 as BMI or for the blood pressure. So, while cleaning the dataset, we replaced the 0 values in the specific column by calculating the median value of the particular column.

4.1.1 Correlation Plot

All the features are positively correlated with "Outcome". From the bar plot given below, we can say that greater Glucose concentration, BMI and Age are the most impacting factors among the 8 parameters for any person to suffer from Diabetes.

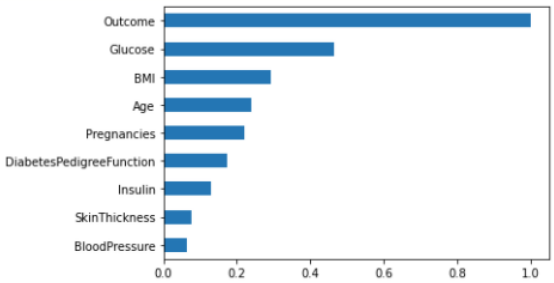

Figure 5: Correlation Plot

4.1.2 Analyzing Outcome vs AGE

From the Diagram below, Age group of 40-60 years are highly vulnerable to suffer from Diabetes.

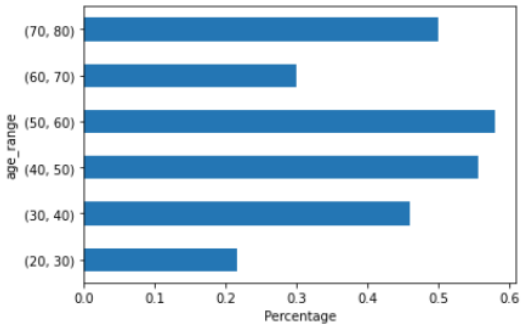

Figure 6: Age vs Outcome

4.2 Classifiers Evaluation

Table 1: Confusion Matrix for Logistic Regression

Actual/Predicted	0	1
0	80	20
1	15	39

Table 2: Confusion Matrix for SVM

Actual/Predicted	0	1
0	81	19
1	14	40

Table 3: Confusion Matrix for Naive Bayes

Actual/Predicted	0	1
0	78	22
1	15	39

Table 4: Confusion Matrix for Decision Tree

Actual/Predicted	0	1
0	77	23
1	15	39

Table 5: Confusion Matrix for Voting without Weights

Actual/Predicted	0	1
0	82	18
1	13	41

4.2.1 Confusion Matrix for Voting with Weights

Table 6: Confusion Matrix for Voting with Weights

Actual/Predicted	0	1
0	83	17
1	12	42

4.3 Overall Classification result for all the Classifiers.

Model	Accuracy	Precision	Recall	F1-score
LR	77.12	0.80	0.90	0.85
SVM	78.35	0.81	0.90	0.85
NB	76.19	0.79	0.88	0.83
DT	75.76	0.84	0.80	0.82
VC	79.59	0.82	0.89	0.85
VCW	81.16	0.82	0.91	0.86

5. Conclusion and Recommendations

In this research work, the performance of the system is enhanced with the combination of multiple classifiers. The voting classification with weights results more accurately than without weight. Here, VCW has high Recall among other classifiers which indicates that class (0 or 1) is correctly recognized i.e. a small number of FN exists. Where, FN explains the cases in which predicted value is False and the actual output is True.

As a part of the recommendation from this research with other sources of data can be trained and tested as well. Where the system can be evaluated with big data so that inter-operability of the model is increased. We can also perform similar hybrid research with similar other machine learning algorithms. We can also try dropping the "Pregnancies" feature from the data set and train the model accordingly and make the model general for all genders.

References

[1] World health organisation. `http://who.int/mediacentre/factsheets/fs312/en/index.html`.

[2] International diabetes federation. `https://www.idf.org`.

[3] Seyed Ataaldin Mahmoudinejad Dezfuli, Seyedeh Razieh Mahmoudinejad Dezfuli, Seyed Vafaaldin Mahmoudinejad Dezfuli, and Younes Kiani. Early diagnosis of diabetes mellitus using data mining and classification techniques. *Jundishapur Journal of Chronic Disease Care*, 8(3), 2019.

[4] Miroslav Marinov, Abu Saleh Mohammad Mosa, Illhoi Yoo, and Suzanne Austin Boren. Data-mining technologies for diabetes: a systematic review. *Journal of diabetes science and technology*, 5(6):1549–1556, 2011.

[5] LHS De Silva, Nandana Pathirage, and TMKK Jinasena. Diabetic prediction system using data mining. 2016.

[6] Bum Ju Lee, Boncho Ku, Jiho Nam, Duong Duc Pham, and Jong Yeol Kim. Prediction of fasting plasma glucose status using anthropometric measures for diagnosing type 2 diabetes. *IEEE journal of biomedical and health informatics*, 18(2):555–561, 2013.

[7] Longfei Han, Senlin Luo, Jianmin Yu, Limin Pan, and Songjing Chen. Rule extraction from support vector machines using ensemble learning approach: an application for diagnosis of diabetes. *IEEE journal of biomedical and health informatics*, 19(2):728–734, 2014.

[8] Nahla Barakat, Andrew P Bradley, and Mohamed Nabil H Barakat. Intelligible support vector machines for diagnosis of diabetes mellitus. *IEEE transactions on information technology in biomedicine*, 14(4):1114–1120, 2010.