# Important Feature Selection & Accuracy Comparisons of Different Machine Learning Models for Early Diabetes Detection

Sajratul Yakin Rubaiat
Computer Science and Engineering Department
Patuakhali Science and Technology University
Patuakhali, Bangladesh
Email: yakin13@cse.pstu.ac.bd

Md Monibor Rahman
Electrical and Computer
Engineering Department
Old Dominion University,USA
Email: mrahm006@odu.edu

Md.Kamrul Hasan
Mathematics, Statistics and
Computer Science Department
Marquette University, USA
Email: mdkamrul.hasan@mu.edu

***Abstract*—This paper proposes an analysis of different methods based on a neural network for predicting type 2 diabetes mellitus (T2DM). The objective of this paper is to find which type of model that works best for predicting diabetes. Pima Indian Diabetes data-set were used in this analysis. The analysis was carried out on this database using two methods. The first method includes Data Recovery followed by feature selection. These features are then used as an input MLP neural network classifier which achieved an accuracy of 85.153%. The second method is based on noise reduction using k-means followed by feature selection. The features thus obtained are used with Random Forest, Logistic Regression and MLP neural network classifier. The maximum accuracy obtained among these classifiers is 77.08%. The consultation shows why Data recovery with MLP is far better than K-means based noise reduction with the different type of classifier.**

## I. INTRODUCTION

Diabetes mellitus[1] is a major public health problem that is approaching epidemic proportions globally.It has particularly increased in the 21 century. Diabetes is caused by several factors, including obesity, consumption of unhealthy food, heredity etc. As of 2015, an estimated 415 million people have been suffering from diabetes worldwide and the trend suggests that the rate will continue to rise. Diabetes has some serious long-term complications including cardiovascular disease, stroke, chronic kidney disease, foot ulcers, and damage to the eyes. For these reasons, researchers need to put more focus on this problem. There are mainly 3 kinds of diabetes. 1. Type 1 DM (caused by failure of the pancreas to produce sufficient insulin) 2. Type 2 DM (the most common cause is excessive body weight and insufficient exercise) 3. Gestational diabetes (occurs in pregnant women with no prior history of diabetes). It should be pointed out that type 2 DM makes up about 90% of the cases.

Data analysis has been successfully applied to various fields of human society, such as weather prognosis, market analysis, engineering diagnosis, and customer relationship management. However, the utilization of disease prediction and medical data analysis still has room for improvement. Every hospital possesses a different kind of basic and medical information, and it is essential to extract useful information from these data to support future medical analysis and diagnosis [2,3]. It is rational to believe that there are several valuable patterns and are waiting for researchers to examine them.

As the number of the patient of diabetes patients is increasing,it is necessary to build a model that can classify patients with high risk of diabetes in the future. Identifying the high risk factors could potentially prevent more cases of diabetes in future.

## II. PIMA INDIAN DIABETES DATABASE

Pima Indian data-set[2] is obtained originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the data-set is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the data-set. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females from Pima Indian Heritage who are at least 21 years old. The information consists of 768 patients (268 instances of 1 and 500 instances of 0) coming from a population near Phoenix, Arizona, USA. 1 and 0 indicates whether the patient is diabetic or not, respectively. Each instance is comprised of 8 attributes, which are all numeric. The data-set consists of several medical predictor variables one target variable and the outcome. Predictor variables are as follow:

- Number of times pregnant (preg)
- Plasma glucose concentration at 2h in an oral glucose tolerance test (plas)
- Diastolic blood pressure (pres)
- Triceps skin fold thickness (skin)
- 2-h serum insulin (insu)
- Body mass index (bmi)
- Diabetes pedigree function (pedi)
- Age (age)
- Class variable (class)

## III. RELATED WORKS

Kamer Kayaer et al.[3] used the PID dataset to evaluate the perceptron-like general regression neural network (GRNN). This study had 576 cases in the training set and 192 cases in the test set. Using 576 training instances, the sensitivity and specificity of their algorithm was 80.21% on the remaining 192 instances. The same number of random training and test sets was used to compare the simulation results. Dilip Kumar Choubey et al.[4] used naive Bayes with the genetic algorithm to evaluate the perception. The accuracy and specificity of their

algorithm was 78.69% on the test set. Manjeevan seera et al.[5] used fuzzy min-max (FMM) neural network to evaluate the model and the accuracy of the algorithm was 78.39%. Hayashi, Y., & Yukita, S.(2016)[6] used Recursive-Rule extraction algorithm with J48graft combined with sampling selection techniques and get 83.83% accuracy.

## IV. Different Section of Analysis

The whole work can be described in 2 ways:

### A. Method 1

This process is done by applying data recovery with the neural network model. This process can be divided into three sections,

- Data recovery.
- Feature selection.
- M.L.P. Classifier.

In first section, data recovery techniques are applied by replacing the missing data with the mean value for making the data set more robust for building a model. After that feature selection is done whose role is to find the features that have most impact on the result. Lastly, a suitable number of hyper-parameters are selected that works well for this data-set.

### B. Method 2

This process is done by applying K-means with different machine learning model.The work can be divided into three sections,

- K-means Algorithm.
- Feature selection.
- M.L.P. Classifier.

The k-means algorithm reduces the noise data and output of the k-means algorithm are used as a feature for the model. The selected features for the model are then tested with different types of the classification method.

## V. Applying Data Recovery with Neural Network Model

### A. Data recovery

Pima Indian dataset contain numerous missing data. Some features like blood pressure or Insulin cannot be zero in a normal person. Number of zeros in the different features are:

- Insulin: 374
- Skin Thickness : 227
- Blood Pressure : 35
- B.M.I : 11
- Glucose : 5

This missing data can affect the results of the model. Model built with this data would be misleading.There are many different methods to recover the data such as:

- Delete the data from the data set (this way it cannot affect the model).
- Replace the missing data with mean value.
- Replace the missing data with the most likely value of this feature.

Pima Indian data set is very small, only 768 samples. Therefore, the model can end up being highly Biased if the training data are deleted. So deleting data is not acceptable. Different person has the different level of insulin level. If we transfer the high number of data with most likely value, we maybe end up with high variance problem. So, the best option is replacing the missing values with the mean value of that particular attribute. At first, replace the missing data with NaN value. NaN is used for replacing the numerical missing value with a string. After that, iterate over the column and find the sum of all the numerical values (NaN is a string value, so it does not add up here). Calculate the general mean by dividing total summation by the number of the entity in the column. Then replace the Nan string with the numeric mean value.

### B. Feature selection

There are many features that does not affect the model. Hence using this kind of feature for training the model will only add up the unnecessary computational power. Fig. 1 contains the Skin thickness graph of all the patient.

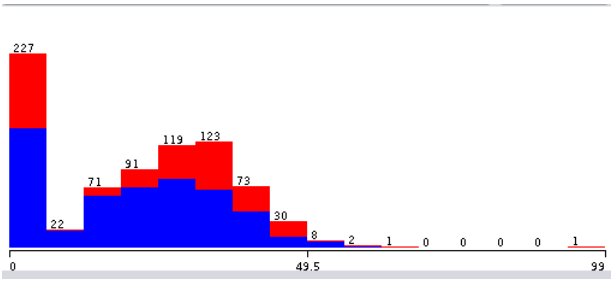


Fig. 1. Skin Thickness graph for "Pima Indian data set". Here X-axis contain skin thickness, Y-axis Contain Number of patient that have diabetes positive(as red) and diabetes negative(as blue).

The first block of the histogram contains a pretty much same number of element from both classes (diabetes positive or diabetes negative). And it still maintains this ratio when the skin thickness is increased. So this cannot be an important feature for the model.

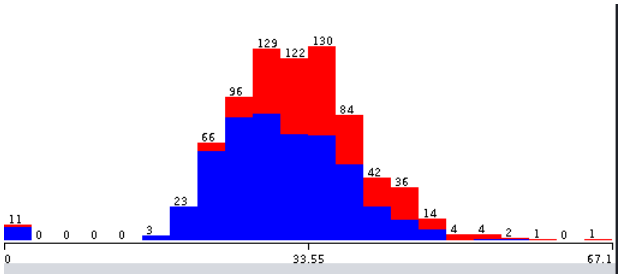


Fig. 2. B.M.I. feature graph from Pima Indian data set(Here X-axis contain B.M.I And Y-axis contain the Number of patient that have diabetes positive(as red) and diabetes negative(as blue).

These data from Fig. 2 and Fig. 3 concludes that, when glucose and B.M.I. levels increase, the risk for diabetes increases significantly. So it provides us with a useful linear property with the B.M.I or glucose level.

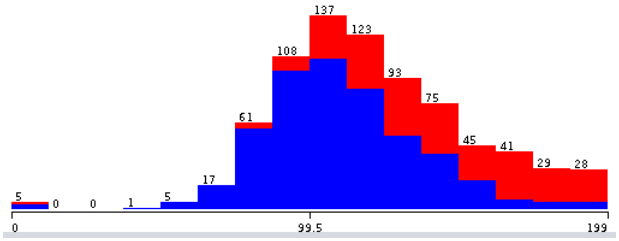


Fig. 3. Glucose level graph for Pima Indian data set(Here X-axis contain glucose level, Y-axis contain Number of patient that have diabetes positive(as red) and diabetes negative(as blue).

Greedy Step wise Search Algorithm is used to select the useful attributes. Greedy Step wise Search Algorithm iterates through each or set of the attribute to calculate which attribute gives the minimum error. The rules for feature selection are as follows:

- Error should never increase.
- Solution should eventually reached.

After evaluating error for the first attribute, the algorithm makes set with the second attribute and calculates the cost. If the error increases then the algorithm just makes the weight zero of the second attribute. After that, it will continue this process for all the others attributes. In the end, as the error cannot increase the solution is eventually found. For calculating the cost function, the algorithm uses cross-validation set for not falling into over fitting or under fitting problem. By analyzing the different graph and investing which feature affects the most, four features were taken. Those features are given below:

- Glucose
- B.M.I.
- Age
- Diabetes Pedigree Function

### C. Multilayer perceptron Classifier

A multilayer perceptron (MLP) is a class of feedforward artificial neural network. We use this algorithm because MLPs are used in research for their ability to solve problems sarcastically, which often allows approximate solutions for extremely complex problems like fitness approximation. There are many hyper-parameters for MLP classifier such as alpha, hidden-layer size, solver, learning-rate decay etc. To find the best model, the different combination of these hyper-parameters are tried randomly and iteratively. Firstly the model gets lower accuracy, it is for high bias problem (because it gives test and training accuracy pretty much same). There are some solution for high bias which are given below:

- Make a bigger network
- Training longer
- Search for different the NN(Neural network) architecture

The appropriate choice is option one and three because the training set is very small so training for longer time is not a very effective process to remove high bias problem. At last, the suitable parameters for our model is obtained which are given below:

- solver=lbfgs
- alpha=1e-5
- hidden layer sizes=(15,7,7,3) (First layer number of node 15,second layer number of node number 7,third layer number 7,fourth layer node number 3,last layer node number 1)

By applying this parameter, the results are as follows:

TABLE I
LOGISTIC REGRESSION ACCURACY

| | |
|---|---|
| **Training Set**: M.L.P. Classifier mean accuracy | 86.733% |
| **Test Set**: M.L.P. Classifier mean accuracy | 85.153% |

## VI. APPLYING K-MEANS WITH DIFFERENT MACHINE LEARNING MODEL

### A. K-means Algorithm

Cluster analysis aims at partitioning the observations into disparate clusters so that observations within the same cluster are more closely related to each other than those assigned to different clusters [5]. Fig. 4 shows the procedure of the K-means algorithm, and the methods for the K-means Cluster algorithm is given below:

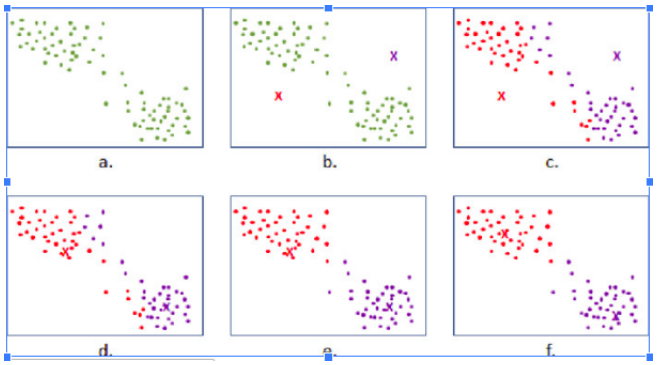


Fig. 4. Visualizing the k-means algorithm for Pima Indian data-set.

- Show all objects (step a). Select K from provided N as the number of initial cluster center (step b). In Fig. 4b, the value of K is 2.
- Calculate distance between each object and cluster center.[6] (step c).
- Recalculate every cluster center to verify whether they are changed.
- Circulate step 2 and step 3 until the new cluster center is the same as the original one, i.e., convergence and end of the algorithm (step e and f).

Fig. 5 shows the output of k-means clustering algorithm applied on Pima Indian Data-set.

K-means algorithm result can be used as an input feature which gives a good advantage in accuracy.

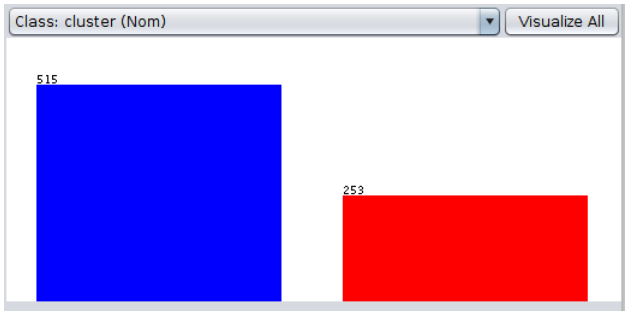


Fig. 5. k-means clustering result in Pima Indian data-set.(Blue for diabetes negative and red for diabetes positive)

### B. Feature Selection

The Greedy step wise search is a feature selection algorithm (Discuss in Section V(B)) that can be used for selecting the useful feature. The greedy step wise search algorithm takes those set of feature that gives minimum error rate. Here are the selected features.

- Pregnancies
- Glucose
- B.M.I.
- Age
- Diabetes Pedigree Function
- Cluster(Output of k-means algorithm)

### C. Classifier

By trying many different kinds of the classifier to examine which method works well. Different classifier like the decision tree, J48, MLP, Logistic regression has different method to evaluate the model. Table 2, 3 and 4 contains the result of the different classifier:

TABLE II
LOGISTIC REGRESSION ACCURACY

| | | |
|---|---|---|
| Correctly Classified Instances | 592 | 77.0833% |
| Incorrectly Classified Instances | 176 | 22.9167% |

TABLE III
M.L.P. CLASSIFIER ACCURACY

| | | |
|---|---|---|
| Correctly Classified Instances | 579 | 75.3906% |
| Incorrectly Classified Instances | 189 | 24.6094% |

TABLE IV
RANDOM FOREST ACCURACY

| | | |
|---|---|---|
| Correctly Classified Instances | 576 | 75% |
| Incorrectly Classified Instances | 192 | 25% |

Logistic regression is most suitable for this kind of problem because its cost function is targeted to make (zero, one) classifier. For this reason, this works well compared to M.L.P. and Random Forest.

## VII. PERFORMANCE EVALUATION

The difference between two proposed methods is that one is using data recovery technique to eliminate noise and other one is using K-means based noise reduction technique. For the given data set, the first method showed improved accuracy because the data recovery with the mean value seems more stable. But the second method which used K-means with neural network did not improved much as the data-set is noisy due to the different kinds of missing value. And therefore applying k-means for clustering the feature has no effect in improving the efficiency of the classifier.

## VIII. DISCUSSION

### A. M.L.P. classifier work

Multilayer Perceptron Classifier is a deep neural network classifier. It cannot be determined in the beginning which hyper-parameter like learning rate, batch size, optimizer, hidden layer size works best for the model. Only after analyzing the errors like high bias or high variance, model can be built by an iterative process that gives a low error rate. In the beginning, M.L.P. classifier starts with hidden layer size=(16,8,2). The model then gives same training set error and test set error. This means it is affected by high bias problem. So it is required to make the model big. After some iterative process and changing different hyper-parameters, an accuracy of 85% is achieved.

### B. Comparison with others experiments

In order to show that this model provided more accuracy of prediction a comparison was made with work done by other researchers using the same data set. Table 5 show the result of different work done by other researchers.

TABLE V
COMPARISON WITH OTHERS EXPERIMENTS

| **Method** | **Accuracy** | **Reference** |
|---|---|---|
| GRNN | 80.21% | Kamer Kayaer[7] |
| Naive Bayes | 78.69% | Dilip Kumar Choubey[8] |
| FMM with neural network | 78.39% | Manjeevan seera[9] |
| J48graft | 83.83% | Hayashi[10] |
| Hybrid model | 84.5% | Humar Kahramanli[11] |
| MLP | 81.9% | Aliza Ahmad[12] |
| ELM | 75.72% | Rojalina Priyadarshini[13] |
| Artificial bee colony | 84.21% | Beloufa[14] |
| Swarm intelligence | 82.03% | Christopher[15] |
| fuzzy rule | 79.37% | Lekkas[16] |
| K-means | 77% | Present study |
| **MLP with Feature Selection** | 85.153% | Present study |

## IX. CONCLUSIONS

This paper presents an analysis of two kinds of the prediction model for diabetes mellitus and making the model adapt to different data-sets. Using these two methods, we can train a model that can predict whether or not someone has diabetes at very early stage with the help of some features like Glucose level, BMI, Age.

The method incorporating k-means based noise reduction technique is easy to implement, but provides lower efficiency and requires more computational power.Whereas, the method that include the data-recovery with neural network requires less computation, but provides higher efficiency of 85%.So, the method that is capable of data-recovery with neural network is more acceptable.

For future work, it is important to bring in hospitals real and latest patients data for continuous training and optimization of our proposed model. The quantity of the data-set should be large enough to train properly and predict with higher efficiency [17,18].

It is more useful and efficient for people to obtain an application about health management of DM on their mobile devices [19, 20].An application can be developed that will provide rational and reasonable health advice to the high-risk group. Diabetes patients can be convinced to use this application to test their blood glucose level, blood pressure, and heart rate.

## ACKNOWLEDGEMENT

We would like to show our gratitude to the Bangladeshi Engineers and Scientists in the USA specially Raihan Masud, S M Iftekharul Alam and Farzana Khalid for mentoring this research project and reviewing our paper as part of their volunteering effort for Ankur International, Portland, USA(https://ankurintl.org/project/about-aus-scholarships/)

## REFERENCES

[1] International diabetes federation (IDF) diabetes atlas (seventh ed.) (2015)
[2] Bellazzi, Riccardo, and Blaz Zupan. "Predictive data mining in clinical medicine: current issues and guidelines." International journal of medical informatics 77.2 (2008): 81-97.
[3] Gittens, Mechelle, et al. "Post-diagnosis management of diabetes through a mobile health consultation application." e-Health Networking, Applications and Services (Healthcom), 2014 IEEE 16th International Conference on. IEEE, 2014.
[4] UC Irvine Machine Learning Repository by "Rexa.info" at the University of Massachusetts Amherst.
[5] G. Guojun, M. Chaoqu, W. Jianhong Data clustering theory algorithm and application (first ed.), ASA-SIAM.M (2007)
[6] Wagstaff, Kiri, et al. "Constrained k-means clustering with background knowledge." ICML. Vol. 1. 2001.
[7] Kayaer, Kamer, and Tulay Yldrm. "Medical diagnosis on Pima Indian diabetes using general regression neural networks."(ICANN/ICONIP). 2003.
[8] Choubey, Dilip Kumar, et al. "Classification of Pima indian diabetes dataset using naive bayes with genetic algorithm as an attribute selection."(ICCCS 2016). 2017.
[9] Seera, M., & Lim, C. P. (2014). A hybrid intelligent system for medical data classification. Expert Systems with Applications, 41(5), 2239-2249.
[10] Hayashi, Y., & Yukita, S. (2016). Rule extraction using Recursive-Rule extraction algorithm with J48graft combined with sampling selection techniques for the diagnosis of type 2 diabetes mellitus in the Pima Indian dataset. Informatics in Medicine Unlocked, 2, 92-104.
[11] Kahramanli, H., & Allahverdi, N. (2008). Design of a hybrid system for the diabetes and heart diseases. Expert systems with applications, 35(1-2), 82-89.
[12] Ahmad, A., Mustapha, A., Zahadi, E. D., Masah, N., & Yahaya, N. Y. (2011). Comparison between Neural Networks against Decision Tree in Improving Prediction Accuracy for Diabetes Mellitus. In Digital Information Processing and Communications (pp. 537-545). Springer, Berlin, Heidelberg.
[13] Priyadarshini, R., Dash, N., & Mishra, R. (2014, February). A Novel approach to predict diabetes mellitus using modified Extreme learning machine. In Electronics and Communication Systems (ICECS), 2014 International Conference on (pp. 1-5). IEEE.
[14] Beloufa, F., & Chikh, M. A. (2013). Design of fuzzy classifier for diabetes disease using Modified Artificial Bee Colony algorithm. Computer methods and programs in biomedicine, 112(1), 92-103.
[15] Christopher, J. J., Nehemiah, H. K., & Kannan, A. (2015). A swarm optimization approach for clinical knowledge mining. Computer methods and programs in biomedicine, 121(3), 137-148.
[16] Lekkas, S., & Mikhailov, L. (2010). Evolving fuzzy medical diagnosis of Pima Indians diabetes and of dermatological diseases. Artificial Intelligence in Medicine, 50(2), 117-126.
[17] Li, Huan, Qi Zhang, and Kejie Lu. "Integrating mobile sensing and social network for personalized health-care application." Proceedings of the 30th Annual ACM Symposium on Applied Computing. ACM, 2015.
[18] Luo, Yan, et al. "GlucoGuide: An Intelligent Type-2 Diabetes Solution Using Data Mining and Mobile Computing." Data Mining Workshop (ICDMW), 2014 IEEE International Conference on. IEEE, 2014.
[19] Schnall, Rebecca, et al. "A user-centered model for designing consumer mobile health (mHealth) applications (apps)." Journal of biomedical informatics 60 (2016): 243-251.
[20] Basar, Md Abul, et al. "A review on diabetes patient lifestyle management using mobile application." proceeding of the 18th International Conference on Computer and Information Technology. 2015.