# Prediction and Diagnosis of Diabetic Retinopathy using Data Mining Technique

Hayrettin Evirgen, Menduh Çerkezi

Sakarya University, Department of Computer Engineering, Serdivan-Sakarya/Turkey

evirgen@sakarya.edu.tr

**Abstract** Diabetic retinopathy is the most common form of eye problem affecting people with diabetes, usually only affects people who have had diabetes for a long time period and can result in blindness. The aim of this study is to examine the naive Bayes algorithm which is one of the classification methods in data mining, and to analyze real life dataset in order to built predictive system for diabetic retinopathy disease. A total of 385 diabetes patients' data were used to train the prediction system. All the categorical features in the dataset were selected by doctors and evaluation was made based on these features. The dataset was obtained at the Eye Clinic of the Sakarya University Educational and Research Hospital. It has been proven with cross-validation that naive Bayes algorithm can be used for diabetic retinopathy prediction with an improved accuracy of 89%.

Keywords: Naive Bayes, Diabetic Retinopathy, Data Mining.

## Introduction

Data mining is the exploration of large datasets to extract hidden and previously unknown patterns, relationships and knowledge that are difficult to detect with traditional statistical methods (Han and Kamber, 2006). The areas where data mining is applied recently include engineering, marketing, healthcare and financial forecasting. Data mining in healthcare is an emerging field of high importance for providing prognosis and a deeper understanding of medical data (Liao and Lee, 2002). The availability of huge amount of patient's data from which to extract useful knowledge, researchers have been using data mining techniques to help health care professionals in diagnosis of diseases. Developing a tool to be embedded in the hospitals management system to help the healthcare professionals in diagnosing patients is important.

The disease predictions play an important role in data mining. Researchers are using data mining techniques in the diagnosis of several diseases such as diabetes (Fang et al., 2009), cancer (Salehi et al., 2010) and heart diseases (Shouman et al., 2012).

In these studies, several data mining techniques are used, such as Naïve Bayes, K-Nearest Neighbor, Decision Tree, Neural Network and also Clustering methods.

Mohammad R. Shakouriet et al. proposed two different techniques from data mining and case-based reasoning. They used K-Nearest Neighbor and Decision Tree-techniques to predict diabetic retinopathy (Shakouriet , 2012). On the other hand, in (Chan et al., 2008) the authors explored the relationship between physiological data and retinopathy using two data mining techniques, namely C5.0 and Neural Network.

In another study, the authors proposed to examine the relationship between the retinal vessel diameter and the risk of retinopathy using the measurement of retinal vessel diameter from fundus photographs (Klein et al., 2004).

This study aims to design a software tool to help health care professionals. At the same time, this application shows the applicability of data mining methods for many problems in the medical field.

## Materials and Methods

*a. Description of Dataset*

The dataset was obtained at the Eye Clinic of the Sakarya University Educational and Research Hospital. The dataset consists of 385 records. In dataset each record consists of 9 features. These are, namely, Glycated Hemoglobin (HbA1C), Hemoglobin (HGB), URE, High-Density Lipoprotein (HDL), Low-Density Lipoprotein (LDL), Diabetes Duration, Triglyceride, Creatine and Glucose. Because Naïve Bayes algorithm does not permit continuous data type, all the values in the dataset are treated as categorical. In Table 1, the diagnosis columns show the categorical values for the corresponding features. The diagnosis column is identified as predictable feature with value "1" for patients with diabetic retinopathy and value "0" for patients with non diabetic retinopathy. All the categorical features in the dataset were selected by doctors and evaluation was made based on these features.

**Table 1**: Demonstrates the clinical feature of the patients in the dataset.

| Feature Number | Description of Feature | Diagnosis (1) | Diagnosis (0) |
|---|---|---|---|
| 1 | Glycated Hemoglobin (HbA1C) | < 6.5 | >= 6.5 |
| 2 | Hemoglobin (HGB) | > 12 | <= 12 |
| 3 | High-Density Lipoprotein (HDL) | > 40 | <= 40 |
| 4 | Low-Density Lipoprotein (LDL) | < 130 | >= 130 |
| 5 | Diabetes Duration | < 5 | >= 5 |
| 6 | Triglyceride | > 150 | <= 150 |
| 7 | Creatine | > 1,2 | <= 1,2 |
| 8 | Glucose | > 140 | <= 140 |
| 9 | URE | > 45 | <= 45 |

*b. Naïve Bayes*

The Bayesian Classification represents a supervised learning as well as a statistical method for classification. Assumes an underlying probabilistic model and it allows us to capture uncertainty about the model in a principled way by determining probabilities of the outcomes. It can solve diagnostic and predictive problems. Naïve Bayes algorithm is based on Bayesian Theorem.

Steps in algorithm are as follows:

1. Each data sample is represented by an *n* dimensional feature vector, X = (X1, X2,…, Xn), depicting measurements made on the sample from *n* attributes, respectively A1, A2, An.
2. Suppose that there are *m* classes, C1, C2,…, Cm. Given an unknown data sample, X (i.e., having no class label), the classifier will predict that X belongs to the class having the highest posterior probability, conditioned if and only if:

   $P(Ci|X) > P(Cj|X)$ for all $i <= j <= m$ and $j != i$

   Thus we maximize P(Ci|X). The class Ci for which P(Ci|X) is maximized is called the maximum posteriori hypothesis. By Bayes theorem,

   $P(Ci|X) = (P(X|Ci)P(Ci))/P(X)$

3. As P(X) is constant for all classes, only P(X|Ci)P(Ci) needs to be maximized. If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, i.e. P(C1) = P(C2) = … = P(Cm), and we would therefore maximize P(X|Ci). Otherwise, we maximize P(X|Ci)P(Ci). Note that the class prior

probabilities may be estimated by P(Ci) = si/s, where si is the number of training samples of class Ci, and s is the total number of training samples.

*c. Cross Validation*

Cross-Validation (CV) is the standard data mining technique for evaluating performance of classification technique. Mainly it's used to evaluate the error rate of a learning technique. In CV a dataset is portioned in *n* folds, where each is used for testing and the remainder is used for training. The procedure of testing and training is repeated *n* times so that each partition of fold is used once for testing.

In a stratified 10-fold Cross-Validation the data is divided randomly into 10 parts in which the class is represented in approximately the same proportions as in full dataset. Each part is held out in turn and the learning scheme trained on the remaining nine-tenths; then its error rate is calculated on the holdout set. The learning procedure is executed a total of 10 times on different training sets, and finally the 10 error rates are averaged to yield an overall error estimate.

*d. Confusion Matrix*

Confusion matrix is a visualization tool which is commonly used to present the accuracy of the classifiers in classification (Han and Kamber, 2006). It is used to show the relationships between outcomes and predicted classes.

The entries in confusion matrix have the following meanings in the context of our study:

- a is the number of correct predictions that an instance is negative,
- b is the number of incorrect predictions that an instance is positive,
- c is the number of incorrect predictions that an instance is negative,
- d is the number of correct predictions that an instance is positive.

**Table 2**: Confusion Matrix

| | | Predicted | |
|---|---|---|---|
| | | Negative | Positive |
| Actual | Negative | a | b |
| | Positive | c | d |

The accuracy (AC) is the proportion of the total number of predictions that were correct. It is determined using equation below:

$$AC = \frac{a+d}{a+b+c+d} \quad (1)$$

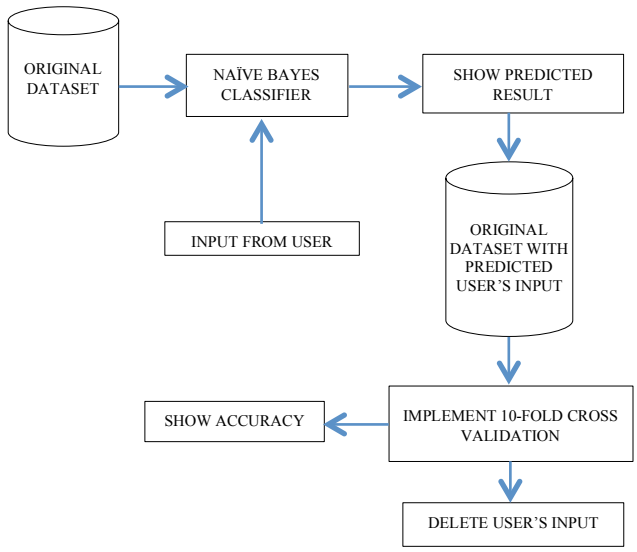


**Figure 1**: Model of Implementation

## Model of Implementation

Fig. 1, depicts the functional block diagram of implementation. Mainly system is working in two phases, in prediction phase and in accuracy evaluation of algorithm. As shown in Fig. 1, the original dataset and input from user is given as input to the classifier. When user enters 9 parameters Naïve Bayes Classifier will predict users' input state, and the prediction will be shown. In the second phase the dataset with predicted user's input will be evaluated with 10-fold cross validation. When 10-fold cross validation finishes the accuracy evaluation the results of evaluation will be shown with confusion matrix. After prediction and accuracy evaluation in the system users' input will be deleted, so the dataset will be original again.

## Application

In this section, implemented software tool for predicting diabetic retinopathy will be introduced. MySQL Database Management System and PHP programming language are used to implement this tool.

The application consists of two parts in the user interface. In these parts users can input patient's record and see the result of prediction. Fig. 2, shows the interface of application. After executing the algorithm, obtained result of disease diagnosis, percentage of accuracy and also some details of test results are shown to users in the result part of user interface. Details of test results contains of 10 confusion matrices with accuracy values obtained from 10-fold cross validation method.

Fig. 2, shows a sample output for a healthy prediction. In this example, the input for 9 variables was entered, and Naïve Bayes predicted with 89.11% accuracy rate.

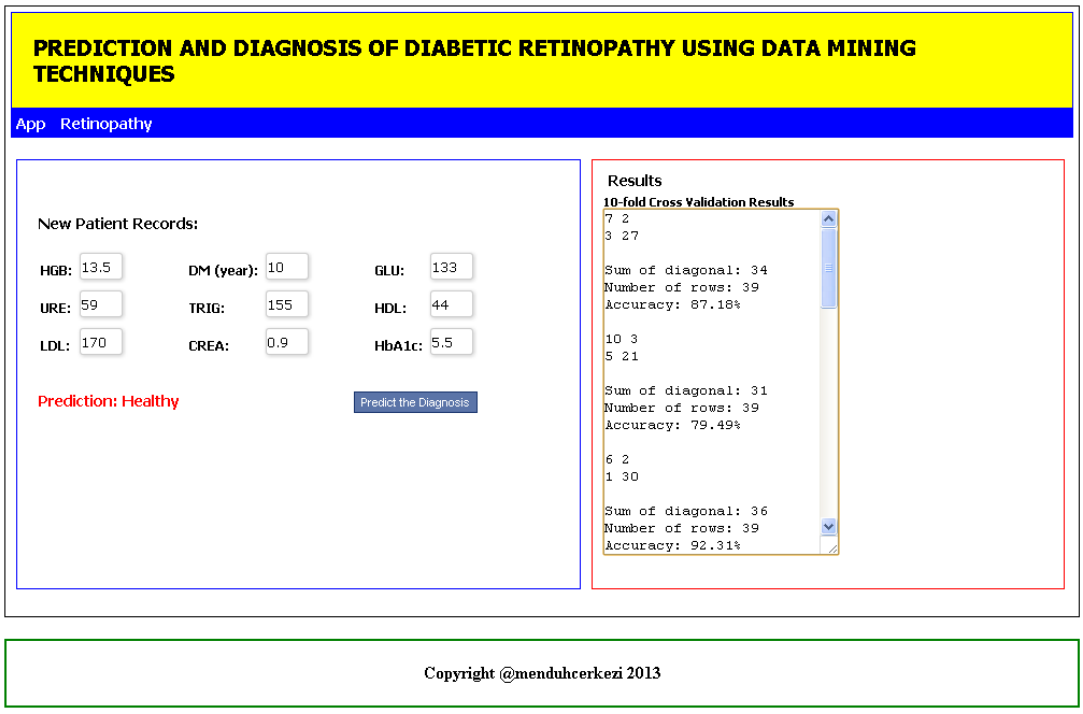


**Figure 2**: Application User Interface

## Results

To obtain and evaluate the test results of Naïve Bayes classifiers, 10-fold cross validation method was used. Hence the dataset is randomly divided into training set and testing set 10 times. Table 3, shows the detail results of 10-fold cross validation. Number of rows column represents testing set of each fold whereas Sum of Diagonal column represents the total number of predictions that were correct. As we mentioned earlier the accuracy is calculated using equation (1).

The result of the accuracy that is obtained is very good using Naïve Bayes algorithm in the real life dataset. The accuracy rate is 89%.

**Table 3**: Detail results of 10-fold cross validation

| Fold | Sum of Diagonal | Number of rows | Accuracy |
|---|---|---|---|
| 1 | 34 | 39 | 87.18% |
| 2 | 31 | 39 | 79.49% |
| 3 | 36 | 39 | 92.31% |
| 4 | 35 | 39 | 89.74% |
| 5 | 33 | 39 | 84.62% |
| 6 | 33 | 39 | 84.62% |
| 7 | 38 | 39 | 97.44% |
| 8 | 38 | 39 | 97.44% |
| 9 | 35 | 39 | 89.74% |
| 10 | 31 | 35 | 88.57% |
| Accuracy: 89.11% | | | |

## Conclusion

This study clearly shows that the results are promising for the application of the data mining techniques into predictions of problem in medical databases.

In this paper, a decision support system was designed for diabetic retinopathy. The system can be served as training tool for medical students. Also, it will be helping hand for doctors. The system can be further enhanced and expanded; it can incorporate other medical features besides in the Table 1, also it can incorporate other data mining techniques. Continuous data can be used instead of just categorical data.

## References


Han, J. & Kamber, M. (2006). *Data Mining Concepts and Techniques*, Morgan Kaufman Publishers.

Liao, S.-C. and Lee, I.-N. (2002). *Appropriate medical data categorization for data mining techniques,* MED. INFORM., Vol. 27, no. 1, 59-67.

Fang, X. (2009). *Are You Becoming a Diabetic? A Data Mining Approach*, Sixth International Conference on Fuzzy Systems and Knowledge Discovery.

Salehi, M., Parandeh N.M., Soltain Sarvestani, A. & Savafi A.A. (2010). *Predictind Breast Cancer Survivability Using Data Mining Techniques*, 2nd Internetional Conference on Software Technology and Engineering (ICSTE).

Shouman, M., Turner, T. & Stocker, R. (2012). *Using Data Mining Techniques in Heart Disease Diagnosis and Treatment*, Japan-Egypt Conference on Electronics, Communication and Computers.

Balakrishnan, V., Shakouri, M. R., Hoodeh, H. & Hakso-Soo, L. (2012). *Predictions Using Data Mining and Case-based Reasoning: A Case Study for Retinopathy*, World Academy of Science and Technology 63.

Klein, R., Klein, B.E.K., Moss, S.E., Wong, T.Y., Hubbard, L., Cruickshanks, K.J. & Palta, M. (2004). *The Relation of Retinal Vessel Caliber to the Incidence and Progression of Diabetic Retinopathy: XIX: The Wisconsin Epidemiologic Study of Diabetic Retinopathy*, Archives of Ophthalmology, vol. 122, pp. 76-83.

Chan, Ch-L., Liu, Y.Ch. & Luo, Sh-H. (2008). *Investigation of Diabetic Microvascular Complications Using Data Mining Techniques*, International Joint Conference on Neural Networks (IJCNN 2008).