Diabetes Prediction Using Classification Algorithms

WARDA FIAZ
Department of computer science
Riphah International University
Lahore, Pakistan
wardafiaz4@gmail.com

KHADIJA TAHIR
Department of computer science
Riphah International University
Lahore, Pakistan
Khadijatahir65@gmail.com

SIDRA RANI
Department of computer science
Riphah International University
Lahore, Pakistan
17mcs1556@gmail.com

MUHHMAD UMAIR ANWAR
Punjab Safe Cities Authority Government Of Punjab
Lahore, Pakistan
m.umair09@live.com

TAYYAB WAQAR
Department of computer science
Riphah International University
Lahore, Pakistan
maliktayyabwaqar@gmail.com

Abstract:

Diabetes and cancer are the leading causes of death in worldwide. Many complications can discover if diabetes will remain untreated. Now machine learning solves it easily. We can use many machine learning algorithms to identify the value of getting for a patient's treatment. Availability and accessibility of data provide us useful

knowledge to apply different techniques. Many machine learning tools are available for prediction and then decide on a patient's treatment. In this paper, the used parameters are Age, Insulin, Pores, Glucose, Blood pressure, and skin thickness. Four different machine learning techniques (K Star, Logistic, and Naive Bayes) are used for comparison.

Keywords: *diabetes, naive Bayes, K star, logistic, classification algorithms, Weka.*

1. INTRODUCTION

Machine learning consists of algorithms that can automatically improve the accuracy of data to apply different experiences. Given data can create and find the patterns or knowledge about predictions for better decision making. Machine learning parts are 1. Training 2. Testing.

We can use machine learning approaches for patient's treatments.

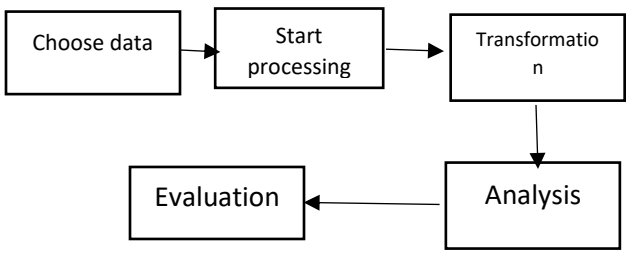
We collect data, transform it, and do analysis on it.

Diabetes can destroy a human's life if people not take treatment. If diabetes will remain untreated, many complications can discover. In diabetes human safer from high blood pressure. Diabetes is a health problem that can disturb the human's blood pressure [1].

In this paper, we used (PIDD) Pima Indian Diabetes Dataset for prediction of diabetes which is the source from the UCI machine learning repository.

We used Weka for analysis. Weka is a useful asset that contains both supervised and unsupervised learning procedures. We are utilizing Weka since it encourages us to assess the consequence of the grouping.

The classification technique is mostly used in the medical and health domain [11]. It provides us information about each step to build the classification model. This paper compares different classification results. Performance is measure by these algorithms. The main objective is to evaluate the algorithms to produce a better prediction.



2. PROBLEM STATEMENT

Diabetes and cancer are the leading causes of death in worldwide. Many complications can discover if diabetes will remain untreated. Now machine learning solves it easily. We can use many machine learning algorithms to identify the value of getting for a patient's treatment**. *T***his is a classification problem of machine learning techniques. The analysis aims to predict whether or not a patient has diabetes or not. And compare different algorithms on a dataset for best accuracy.

Questions:

- Which algorithm's accuracy is best?
- Analysis for prediction?
- Significance of Weka?

3. RELATED WORK

The main objective of this research paper ("Diabetes prediction using classification algorithms") is to describe the classification algorithms and prediction of diabetes. Many classification techniques are used for prediction. This paper, analysis of diabetes prediction and evaluate the best accuracy for decision making. You can also apply clustering, classification for data. But the main aim of this paper is that how we can save human life through the prediction of diabetes.

Mr. Chintan Shah, [1] explains the models that build on based on classification algorithms.

C.M Velu [2] present visualizing of the classification of diabetes using the Weka tool.

Naveed Kishore G, V. Rajesh, K. Sumedh, A. Vasi, [10] present the classification of diabetes using machine learning algorithms and evaluate the performance of prediction.

Depti Sisodia, [11] evaluate the diabetes dataset classification using through prediction using machine learning.

Chitra jigan, [12] presents the prediction by using the SVM classifier and evaluate its performance for decision making about patient's treatment.

M. Venka Das, [4] applied classification techniques for the subtype of lung cancers by using tool Weka. He used a decision tree for prediction.

S. Suru [13] applied machine learning techniques on PID (PIMA INDIAN DIABETES) Dataset to find out the classification accuracy by prediction.

Tajas N Josi [14] compare the machine learning techniques, supervised and unsupervised on diabetes dataset to help out the better treatment of diabetes to save human life.

4. DATASET AND METHODOLOGY

For the evaluation of the result, we collect data and build a model process.

- *Data collection*

Data collection is information about the dataset. The dataset for analysis is collected from the Kaggle UCI machine learning repository. The dataset is taken from the national institute of Diabetes and Digestive and kidney diseases. The main objective of collected data is found out a patient whether or not he has diabetes. The data is received on 9 May 1990. Vincent Sigillito is the donor of this database. Dataset has 9 attributes. Fig [2] [3] [4]. No missing value in it. The number of instances 786. Attribute 8 plus.

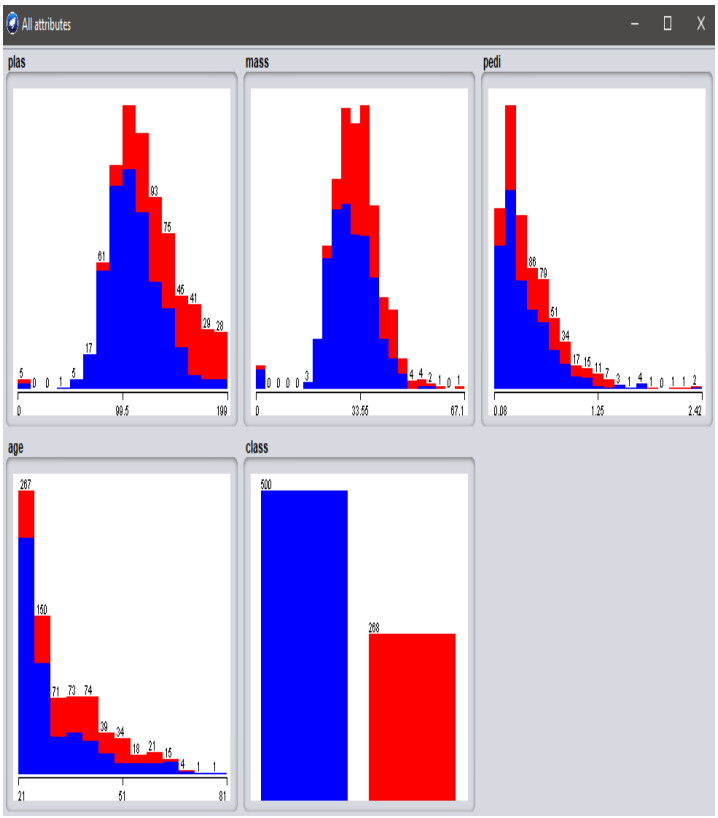


Figure 1 attributes

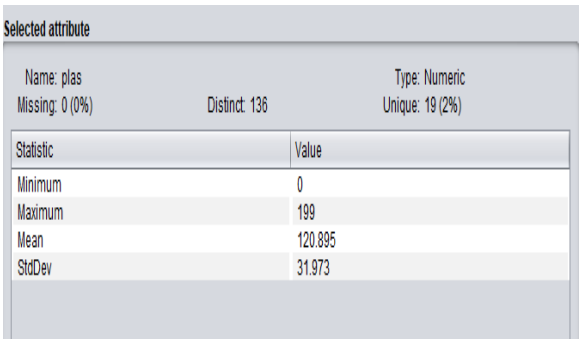


Figure 2 selection of attributes

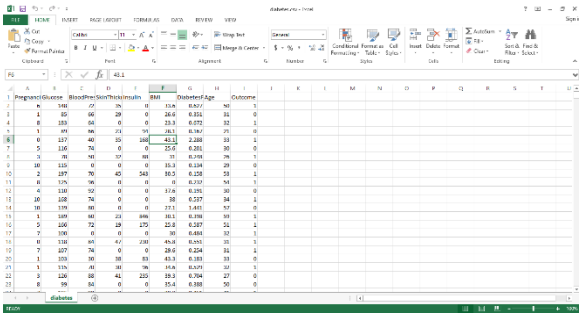
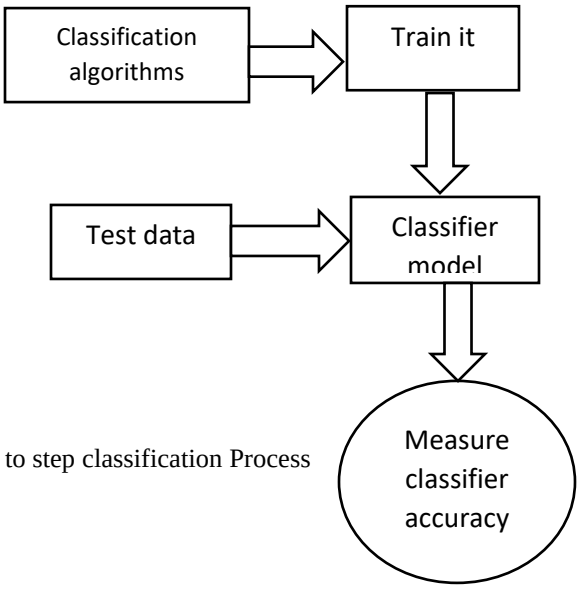


Figure 3 dataset

- *Classifier process*

For step by step classification process, it's important to sort critical diseases to make them target variable or class tags for labeled learning techniques. The step by step classification process is shown in the blow.

Weka is used for this purpose because it's the easiest tool for the evaluation of an algorithm's accuracy. It provides us with supervised as well as unsupervised learning algorithms. So, we can use both.



Step to step classification Process

5. USED ALGORITHMS

In this paper, we are using the following algorithms:

A. The Naive Bayes classifier

A Naive Bayes classifier is a collection of many algorithms based on Bayes' theorem. It's not a single algorithm it depends on all family where all share common principles e.g. pair of features that are classified independently.

Naive Bayes theorem is:

$$P(y|X) = p(X|y)p(y)\backslash p(X)$$

To build a model use Weka and find out the accuracy of the model we applied the cross-validation is 10 fold as a test option.

We get the following information:

```
Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances         182               79.1304 %
Incorrectly Classified Instances        48               20.8696 %
Kappa statistic                          0.4957
Mean absolute error                      0.2893
Root mean squared error                  0.3814
Relative absolute error                 64.3725 %
Root relative squared error             81.7435 %
Total Number of Instances              230

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   MCC      ROC Area  PRC Area  Class
                 0.880     0.403     0.827     0.880     0.853     0.498     0.840     0.904     tested_negative
                 0.597     0.120     0.694     0.597     0.642     0.498     0.840     0.734     tested_positive
Weighted Avg.    0.791     0.314     0.785     0.791     0.787     0.498     0.840     0.851

=== Confusion Matrix ===

   a   b   <-- classified as
 139  19 |   a = tested_negative
  29  43 |   b = tested_positive
```

Figure 4 naïve Bayes

Naive Bayes gives an accuracy of 79%.

In the result total 230 instances, 182 are correctly classified. 48 are incorrectly classified.

B. K star classifier

It is an instance-based classifier that is the category of a test instance that relies upon the category of those training instances almost prefer it, as determined by some similarity function. It's different from learning instance-based.

```
Time taken to test model on test split: 0.63 seconds

=== Summary ===

Correctly Classified Instances         170               73.913  %
Incorrectly Classified Instances        60               26.087  %
Kappa statistic                          0.3794
Mean absolute error                      0.3103
Root mean squared error                  0.4302
Relative absolute error                 69.0444 %
Root relative squared error             92.2047 %
Total Number of Instances              230

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.829    0.458    0.799      0.829    0.814      0.380    0.777     0.886     tested_negative
               0.542    0.171    0.591      0.542    0.565      0.380    0.777     0.571     tested_positive
Weighted Avg.  0.739    0.368    0.734      0.739    0.736      0.380    0.777     0.787

=== Confusion Matrix ===

   a   b   <-- classified as
 131  27 |   a = tested_negative
  33  39 |   b = tested_positive
```

Figure 5 k star

It gives an accuracy of 73%.

In the result total 230 instances, 170 correctly classified. 60 are incorrectly classified.

C. Logistic classifier

Logistic is a powerful classifier that directly predicts the probabilities through something called logit transform. We can easily measure accuracy and built a class through is. It is a statistical model that uses binary but dependent variables.

After experiment logistic we get the following information:

```
Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances         185               80.4348 %
Incorrectly Classified Instances        45               19.5652 %
Kappa statistic                          0.5216
Mean absolute error                      0.2987
Root mean squared error                  0.3748
Relative absolute error                 66.4494 %
Root relative squared error             80.3159 %
Total Number of Instances              230

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.899    0.403    0.830      0.899    0.863      0.527    0.848     0.901     tested_negative
               0.597    0.101    0.729      0.597    0.656      0.527    0.848     0.767     tested_positive
Weighted Avg.  0.804    0.308    0.799      0.804    0.799      0.527    0.848     0.859

=== Confusion Matrix ===

   a   b   <-- classified as
 142  16 |   a = tested_negative
  29  43 |   b = tested_positive
```

Figure 6 logistic

We get 80% accuracy by the logistic classifier.

In the result total 230 instances, 185 are correctly classified. 45 are incorrectly classified.

6. RESULTS:

Logistic classifier provides us the high accuracy than others. It provides us 80% accuracy. We can see below:

Table 1 Supervised Simulation Errors

Instance of Algorithms	The error of Root Mean sq.	The error of Mean absolute	The error of Relative absolute	The error of Root relative sq.
Naive Bayes	0.3814	0.2893	64%	81%
K star	0.4302	0.3103	69%	92%
logistic	0.3748	0.2987	66%	80

Table 2 Weka Evaluation Criteria

Classification total instance=230	The instance that Correctly classified	The instance that Incorrectly classified	Statistic Kappa	The time that a model take to build
Naive Bayes	182(79%)	48(20%)	0.4957	0 secs
K star	170(73%)	60(26%)	0.3794	0.63 sec
Logistic	185(80%)	45(19%)	0.5216	0 secs

7. CONCLUSION AND FUTURE WORK

This work performs by Weka using 3 classification machine learning algorithms. We compare the result of algorithms in terms of accuracy and time building model. Weka is a powerful and efficient tool for the evaluation algorithm's results. We have done a prediction on the diabetes dataset. This work shows us that logistics is best than the other two machine learning algorithms. In the future, we will use a combination of datasets and compare the accuracy of

all algorithms. Further, we will add more algorithms for comparison and evaluate the performance of every dataset.

8. REFERENCES

[1] Mr. Chintan Shah, Dr. Anjali G. Jivani, "Comparison of Data Mining Classification Algorithms for Breast Cancer Prediction", IEEE, DOI:10.1109/ICCCNT.2013.6726477.

[2] R. ROBU and C. HORA, "Medical data mining with extended WEKA", IEEE, DOI: 10.1109/INES.2012.6249857.

[3] C. M. Velu and K. R. Kashwan," Visual Data Mining Techniques for Classification of Diabetic Patients", IEEE, DOI: 10.1109/IAdCC.2013.6514375.

[4] Kanika Chuchra and Amit Chhabra," Evaluating the Performance of Tree-based Classifiers Using Ebola Virus Dataset", IEEE, DOI: 10.1109/NGCT.2015.7375168.

[5] M.Venkat Dass, Mohammed Abdul Rasheed, Mohammed Mahmood Ali," Classification of Lung cancer subtypes by Data Mining technique", IEEE, 2014, DOI: 10.1109/CIEC.2014.6959151.

[6] SP.Chokkalingam and K.Komathy," Comparison of Different Classifiers in WEKA for Rheumatoid Arthritis", IEEE, DOI: 10.1109/ICHCIIEEE.2013.6887821.

[7] https://www.google.com/search?q=what+is+naive+bayes+algorithm&rlz=1C1CHBD_enPK814PK817&oq=what+is+naive&aqs=chrome.2.69i57j0l7.8428j1j7&sourceid=chrome&ie=UTF-8

[8] https://www.google.com/search?rlz=1C1CHBD_enPK814PK817&sxsrf=ALeKk02xOyeTpRLCrNjV3m5a6cc29P-hcA%3A1588572153490&ei=-a-vXtKlHcWVsAf59KfoDQ&q=k+star+classifier&oq=k+star&gs_lcp=CgZwc3ktYWIQAxgAMgQIIxAnMgQIABBDMgIIADICCAAyBAgAEEMyBAgAEEMyAggAMgIIADICCAAyAggAOgcIIxDqAhAnOgUIABCDAToFCAAQkQJQpu0HWLL1B2CchQhoAnAAeACAAcwCiAGUDpIBBTItMy4zmAEAoAEBqgEHZ3dzLXdperABCg&sclient=psy-ab

[9] Barakat, et al. "Intelligible Support Vector Machines for diagnosis of Diabetes Mellitus." IEEE Transactions on Information Technology in Biomedicine, 2009.

[10] Naveed Kishore G, V. Rajesh, "prediction of diabetes using machine learning classification algorithms" International journal of scientific and technology research volume 9, issue 01, January 2020.

[11] Deepti Sosodia, Dilip Singh Sisodia, "prediction of diabetes using classification algorithms" International Conference on Computational Intelligence and Data Science (ICCDS 2018).

[12] Chitra jegan, "Classification of diabetes diseases using Support Vector Machine" International Journal of Engineering Research and Applications (IJERA) vol. 3 issue 2, March-April 2013.

[13] S. Saru, S. Subashree "Analysis and prediction of diabetes using machine learning" International Journal of Emerging Technology and Innovative Engineering" volume 5, issue 4 April 2019.

[14] Tejas N. Joshi, prof. Pramila M. Chawan, "Diabetes prediction using machine learning techniques" International Journal of Engineering Research and Application, vol. 8 issue January 2018.

[15]https://www.google.com/search?rlz=1C1CHBD_enPK814PK817&sxsrf=ALeKk00h-_e7G2qTVp5FKWuYaGB_4sOo3g%3A1588503287640&ei=96KuXv_fJuLDxgOZzYHoBA&q=whta+is+machine+learning+techniques&oq=whta+is+machine+learning+techniques&gs_lcp=CgZwc3ktYWIQAzIECAAQDTIICAAQCBANEB4yCAgAEAgQDRAeMggIABAIEA0QHjoECAAQRzoGCAAQDRAKUIomWMMzYKY2aABwAngAgAG2AogBrxGSAQUyLTYuMpgBAKABAaoBB2d3cy13aXo&sclient=psy-ab&ved=0ahUKEwi_t7zxw5fpAhXioXEKHZlmAE0Q4dUDCAw&uact=5