Diabetes disease prediction using decision tree for feature selection

To cite this article: Jayakumar Sadhasivam et al 2021 J. Phys.: Conf. Ser. 1964 062116

View the article online for updates and enhancements.
Diabetes disease prediction using decision tree for feature selection

Jayakumar Sadhasivam¹, V Muthukumaran², J Thimmia Raja³, Rose Bindu Joseph⁴, Meram Munirathanam⁵ and J M Balajee ⁶

¹School of Information Technology and Engineering (SITE), Vellore Institute of Technology, Vellore, India.
²Department of Mathematics, School of Applied Science, REVA University, Bengaluru – 64.
³School of Computing, Kalasalingam Academy of Research and Education, Tamilnadu-626128, India.
⁴Department of Mathematics, Christ Academy Institute for Advance Studies.
⁵Department of Mathematics, Rajiv Gandhi University of Knowledge Technologies.
⁶School of Information Technology and Engineering, Vellore Institute of Technology, Vellore, India.
E-mail: *jayakumarsvit@gmail.com

Abstract. In this paper more than one approaches are evaluated to optimise machine learning models for diabetes disease diagnosis. The main goal is to sort the medical data computation and choose the most suitable parameters to construct a faster and more perfect model using feature selection. Reducing the number of features to construct a model could direct to more useful machine learning algorithms which helps the doctors to focus on what are the most significant assessment to take into story. Feature selection is one of the process in machine learning which choose a subset of topical features namely variables for construction of models. In this research paper we use three feature selection techniques like Recursive Feature Elimination (RFE), Genetic Algorithm (GA) and Burota Package. After using feature selection at the end we use Decision Tree to predict the diagnosis Diabetes using a dataset named Pima Indian Diabetes Dataset and verify the performance of result model.

1. Introduction

Diabetes is a disease which occurs when the glucose or sugar level in a body of a person increases abruptly. If the insulin production of the body is not good then what happens is the blood sugar tends to increase and there is an adverse effect on cells of the body because they are not able to respond correctly as they should to the insulin. When a person eats food that time the food gets converted to energy and as a result of which blood sugar releases in the body. If the blood sugar increases then a person will experience immediate urination and they become hungry. The pancreatic glands in our body produces insulin which helps glucose to inject to our body cells. Diabetes if not treated in time can lead to adverse consequences and can even lead to death so it is very important to prevent it well in advance so regular check-up is required in order to know what is the condition of the body and if there is a chance of occurrence of diabetes then what are the necessary steps that have to be taken in order to prevent it.

Machine learning on the other hand is a technique by which systems can make intelligent decisions based on some algorithms which are fed into them. In this paper also we are applying machine
learning algorithms to predict diabetes disease in a patient. As machine learning incorporates certain algorithms which help us to find or predict certain outcomes in this case also it will help us to find the diabetes disease occurrence in a patient based on certain attributes which are there in the dataset that we have taken. We are focussing more on applying certain feature selection techniques to select the best attributes in our dataset in order to predict better and efficient result out of the dataset. Some important feature selection techniques that we are going to apply to our dataset will be namely Recursive Feature Selection (RFE) technique, Boruta package and Genetic algorithm. These feature selection techniques we will be able to find the best attributes which are there in our dataset and we can also omit the attributes which are not that important. If the performance is to be measured then, the Decision tree algorithm is used to predict the diagnosis of Diabetes.

Feature selection and Dimensionality reduction (DR) are both different things. If we use both then they will decrease the total number of features from the given dataset but with the DR technique what we do is we perform the task which is same but just we make some advanced union of features. This paper gone through some of the existing literatures on this particular disease diagnosis, features are present in Pima Indian Diabetes dataset (which is the main database for our paper) and the local dataset gathered from the local hospitals, working of various feature selection techniques, and evaluate the results the Decision tree classifier of machine learning for performance and accuracy of the model.

2. Literature Survey

Hasan Akan et al., [1] addressed those different feature selection technique namely “Principal Component Analysis” & “Sequential forward Selection on diabetes disease dataset” & verified the accuracy & performance on the decision tree classifier & Bayesian. The performance of classifier obtained from the feature reduction techniques are compared. This study produces the importance of the feature reduction techniques & they showed that the principal component analysis reduces the number of features more & gives the accurate performance of the model with decision tree classifier as the suitable classifier for such techniques.

Mr.AkhilJabbbar et al., [2] described the risk-score of heart disease prediction using algorithm. Association rules mining has been generated using various feature selection techniques to extract the features which adds high importance to the accuracy & performance. They used classification technique with association to increase the performance. It is appropriate for the utilization in the field of medical. These obtained rules will help the doctors to check the disease issues of patients.

Sadhanaa Tiwari &Birmohan Singh [3], had labelle a innovative technique of selecting features in order to increase the accuracy of performance. They had verify the performance of their work on heart-data & sonar-dataset. The proposed technique which retrieving a important set of features by using ReliefF method used for ranking & a set of unnecessary attributes are extracted using SBS. ReliefF is a attribute which ranking method in which it ranks the attribute based on the importance & their mutual relationship. SBS also holds on this correlation technique of selecting features which starts selecting features from the bucket & remove after it gets selected. They conclude that the accuracy they get with the combined approach of “ReliefF& SBS” is more comparing to the individual.

Abu BakrSoliman et al., [4] proposed different methodologies to build learning model for the early detection of the “Parkinson” disease. For feature selection especially they compare the wrapper and Filter techniques. After that they selected a good learning algorithm to predict which technic might help them to measure performance and accuracy for each technic. They concluded that using filter method it could not provide most related subsets as it measures each feature as a single factor. On the immediate other side, the wrapper method gives the intimating subset of related features, but it could put out a lot of time.
T S Suganya and Dr S Murugavalli [5] used the “Shape and Hough method” of feature extraction and “Group Search Optimization (GSO)” and “Firefly algorithm (FA)” for feature selection to recognize the ancient of Tamil script. They verified the accuracy on neural network, “Naïve Bayes” and “KNN classifier”. “Firefly Algorithm (FA)” comes from biochemical, social aspects and behaviour of animal searching is the basis for GSO which is purely a nature-inspired algorithm.

Bin Wei et al., [6] presented an algorithm with a cross feature selection technique called F score & Compact Genetic Algorithm which rely on SVM. This technique merge the advantages of wrapper & filter techniques of feature selection, by which it remove the useless attribute & lessen the time of computing. They experimented the proposed technique to predict the lung maycer disease using dataset. To verify the performance of an algorithm that they proposed, they verified it with Naive Bayes classifier. They conclude that the proposed algorithm receives better accuracy compared to other implemented techniques.

V. Muthukumaran and D. Ezhilmaran. [7] aims to improve the accuracy and performance of the existing diagnostic method for Type 2 Diabetes using machine learning. Actually they proposed an algorithm that choose the very much essential attributes from the Pima Indian Diabetes Dataset with Goldberg’s Genetic algorithm in the feature reduction stage and to check the accuracy and performance they used the “Multi Objective Evolutionary Fuzzy Classifier” on the dataset. The algorithm uses the principle of minimum rules and maximum classifier rate. Using the feature selection with GA the total number of features is reduced to 4 from 8 and the rate was improved to 83.0435 % with NSGA II.

Nichole Challita et al., [8] presents a very advanced technique of feature selection with the help of neural network. The new algorithm highlights the best feature among the existing ones. The “weighting-based” technique of the input attributes is used in neural network to select the best feature. The aim of this method is to change the weights of the different features to minimize the error of classification. They found that the attributes with high weights are vital and should be select for good result. The performances that they obtained using this method select the best feature on spurious dataset.

Mohammed Siyad B et al., [9] have presented a genetic based evolutionary approach for selecting the appropriate features to improve the accuracy. The proposed approach works in two stages. First stage contains the three available feature selection techniques and is applied individually and the selected features are fused together. Second stage uses Genetic Algorithm to extract the best features and at the end the model is trained using Random forest classifier. An important observation from their analysis is that the optimization using genetic algorithm obtains best prediction accuracy compare to the existing ones. Hence their proposed fusion based on feature selection can be applicable for fine disease diagnosis.

Ganesh Gopal Deverajan et al. [10] proposed an approach to select features before classification technique in order to improve the performance of that particular models. For feature selection, SVM-RFE (Reverse Forward Elimination) and gain ratio algorithms are applied which assigns the value to each feature. This approach is actually support to increase accuracy and reduce executional time. Random Forest and Naïve Bayes classifiers are used to predict the accuracy for the generated models. Experimental results shows that proposed approach that select the feature increases accuracy for both the models.
3. Proposed Work

Machine learning has emerged to be the latest trend in the world nowadays. Machine learning is extensively used in health care systems nowadays because it makes use of the large health records. We have used machine learning because it is considered to be the best approach for our objective of prediction of the diabetes disease in a patient. In this section the data pre-processing steps are being discussed and the dataset is also described in a tabular format. The fig 1 here represents the working nature of this paper. It compares between different feature selection techniques as well as their performance and accuracy.

![Figure 1. Work Flow Diagram](image)

- **Dataset**
The dataset used here is the pimaindian diabetes disease dataset obtained from the UCI repository. This dataset contains 768 entries with 8 numerical features. The attributes mentioned in the dataset are as follows:
Pregnancies, Glucose, Blood pressure, Skin thickness, Insulin, BMI, DPF, Age and Outcome.
The Outcome is the target attribute in the Pima Indian dataset.
The Local Dataset collected from hospital for Diabetes has used for prediction the diabetes disease which contains 367 entries with 17 features. Attributes present in the dataset are mentioned below:
Cholesterol, stabilized glucose, hdl, ratio, glyhb, location, age, gender, height, weight, frame, bp1s, bp1d, waist, hip, time-ppm and waist/hip. Glyhb is the target attribute in the local dataset.

- **Feature Selection Techniques**
Feature selection techniques are used because it always helps us to find the best attributes required to predict the outcomes with utmost efficiency and by this the performance of the model also enhances. This also solves the problem of overfitting. There are some important feature selection techniques which we are using like Recursive Feature elimination (RFE), Genetic algorithm and Burota package

3.1. **Recursive Feature Elimination Selection (RFE)**
In this technique some part of the features are taken into consideration and the model learns by making use of this features. The information that we get from the model which has been trained based on that it again adds or removes features. This is a greedy approach where the objective is to get all the important features at one place. It also generates a copy and the good or bad features which are there are separated during each turn. The features which are left out are developed until all the features get completed. Then the rank of all attributes are given.
Fig 2. Describes the basic process involved in the Recursive Feature Elimination selection technique.

![Recursive Feature Elimination Process](image1)

**Figure 2.** Basic flow of the Recursive Feature Selection technique

4. **Genetic Algorithm (GA)**

If a certain problem is given and we are computing the nearest solution required for the problem then we do so by using Genetic algorithm. The objective is to reduce the population of a single attribute with the help of some set. Then the fitness function is calculated. The fittest single attribute is selected using such techniques. The single attributes which come next have genotypes which results in parental form. In the next iteration the single attribute which is generated are having genotypes resulting in parental form.

![Genetic Algorithm Flowchart](image2)

**Figure 3.** Basic flow of the steps involved in Genetic Algorithm Feature Selection

**Initialization**

Initially from the population all the single attributes are obtained. The solutions which are achieved are included in the populations and also the size of the population depends on identification pf the problem. The population covers the whole area of the possible solutions.
Selection
When the iteration of the algorithm takes place then it needs some portion of the previous iteration and by doing so will generate the algorithm. How we select features entirely depends on the validation function where based on certain fitness parameters. As the best fitness function is taken into consideration so these algorithms are generally the boosting based ones. Sometimes some sample or a part of data is used as a sample and fitness of model being generated. This selection process takes a bit time.

Reproduction
In genetic algorithm the most important stage is reproduction where some population is generated through a rigorous process of operations which are known as “crossovers” and “mutations”. The new population which gets created is called as the parent population. There can be one or more parent populations when new population is generated and this is known as child population. With the help of crossover and mutation the child process is prepared and if some small changes is done otherwise it’s the same as parent.

Crossover
Here if we swap or cross the patterns formed by the parent process then a new population is generated. Whether the population generated is closer to validation function requirements or not is tested in crossover.

Mutation
In a particular population the bits need to be changed by converting a bit 0 to 1 or a pattern change to improve all the weights and fitness of the population.

1. A Genetic Algorithm procedure:
   i. The starting population needs to be selected randomly.
   ii. The fitness function is calculated for the population. I
   iii. Iteration of the group follows till it ends.
   iv. The best fit from the population is chosen.
   v. Mutation and crossover techniques are generated
   vi. The individual fitness function of the single attribute is generated.
   vii. Substitute the less fitted population with the recently generated single.

Boruta Package
When relationship between multiple variables is to be considered then Burota package works best. It incorporated random forest and is actually an extended version of it for feature selection. When all attributes in a dataset have some relationship with the target variable then it is used. It is optimal compared to other feature selection techniques. We can also call it a wrapper method which uses Random forest.

The Algorithm for Boruta package.

1. First the duplicate copies of all attributes are created which are known as the shadow copies.
2. Then the efficiency of both the actual attributes and the shadow attributes are taken into consideration and this also calculates the importance of each and every attribute.
3. In each time it will verify the Z score margin and eliminate the attribute which is not at all important.
4. At last all the good attributes are covered like this and all bad attributes are rejected as well as each attribute in the dataset is covered

**Decision tree classification algorithms**

Decision tree is a supervised learning algorithm where each internal node in a tree represents the attribute in a dataset and each leaf node present denotes the target variable or the outcome. Decision tree is extensively used algorithm for machine learning projects because it is very easily understood and it is not so complex. It is used most of the times for classification and regression problems. The working nature of decision tree is as follows:-

1. Firstly all the samples are at the top main node
2. Features can be categorical in nature
3. The best splitting attribute is selected and then the model is created based on that
4. A statistical approach is used to build the model and applied on the test data

![Decision Tree Model](image.png)

**Figure 4. Decision tree model**

5. **Results and discussion**

We have used R programming language to implement the project which mainly uses statistical and graphics computation. Many types of algorithms are there in R like Regression, Classification and Clustering

5.1. **Confusion matrix**

Confusion matrix is a type of metrics used for target variables which are generated for the local dataset which actually is in a R * R format where R represents the target variable. Confusion matrix even shows the correct and incorrect specified instances

- True Positive (X1): Predicted values is Yes and Actual value is also Yes.
- True Negative (X2): Predicted values is Yes and Actual value is No.
- False Positive (X3): Predicted values is No and Actual value is Yes.
- False Negative (X4): Predicted values is No and Actual value is No.

The Table 1 shows the confusion matrix obtained using various feature selection techniques.
Table 1. Confusion matrix obtained using rfe, boruta and ga

|               | RFE         | Boruta     | GA         |
|---------------|-------------|------------|------------|
|                | Predicted   | Predicted  | Predicted  | Predicted | Predicted |
|                | Yes         | Yes        | Yes        | No        | No        |
| Actual Yes     | 90          | 90         | 96         | 1         | 90        |
| Actual No      | 6           | 6          | 8          | 8         | 6         |

5.2. Results of Feature Selection

5.2.1. Features extracted using RFE
The input for the RFE is shown in Fig 3 which consists of 9 attributes from Pima Indian Diabetes.

```
"Pregnancies", "Glucose", "BloodPressure",
"SkinThickness", "Insulin", "BMI",
"DiabetesPedigreeFunction", "Age", "Outcome"
```

Figure 5. Input given to the feature selection RFE algorithm

The output obtained from RFE is shown in Fig 4 which consists of 5 attributes from Pima Indian Diabetes.

```
The top 5 variables (out of 8):
Glucose, BMI, Age, Pregnancies, SkinThickness
```

Figure 6. Output given by the feature selection RFE algorithm

5.2.2. Features extracted using genetic algorithm
The inputs are the same as they were for RFE algorithm

```
"SkinThickness", "BMI", "Outcome"
```

Figure 7. Output given by the feature selection genetic algorithm

5.2.3. Features extracted using Boruta package

```
"Pregnancies", "Glucose", "SkinThickness",
"Insulin", "BMI", "DiabetesPedigreeFunction",
"Age", "Outcome"
```

Figure 8. Output given by the feature selection Boruta package

5.3. Performance metrics used
Accuracy is the ratio of correctly identified samples to the given test samples.

Sensitivity is the measures of accurately recognized favourable.
Specificity is the measures of accurately recognized unfavourable.

**Accuracy**

| Table 2. Accuracy obtained without feature selection |
|---------------------------------|-----------------|
| **Dataset**                    | **Accuracy (Decision tree)** |
| Pima Indian Diabetes Dataset   | 74.48 %          |

| Table 3. Accuracy obtained with feature selection for pima diabetes dataset |
|---------------------------------|-----------------|
| **Feature Selection techniques** | **Accuracy (Decision tree)** |
| Pima Dataset                    |                 |
| Recursive Feature Elimination   | 66.53 %          |
| Bu Package                      | 70.71 %          |
| Genetic Algorithm               | 63.18 %          |

5.4. Visualization

The Fig. 9 shows the accuracy keep with and without feature selection techniques. The Fig. 9 shows the accuracy in terms of bar chart with feature selection using the Pima Diabetes Dataset and Fig. 9 shows the accuracy obtained in bar chart with feature selection techniques (RFE, Burota, GA) using Local Diabetes Dataset.

The Fig. 9 shows the accuracy obtained without using Feature Selection techniques for Pima Diabetes dataset and Local Diabetes dataset.

![Accuracy for Pima Diabetes Dataset](image)

**Figure 9**. Accuracy obtained with feature selection by using Pima dataset

6. Conclusion

Feature selection is vital for the information to be prepared. In this paper, we take three element choice strategies to be specific Recursive Feature Selection, Feature determination utilizing Burota Package and the Genetic Algorithm utilized in Feature selection. By applying every one of these techniques on
diabetes dataset utilizing choice tree ID3 calculation, we acquired that by applying Feature selection calculation on standard dataset, in particular, Pima Indian Diabetes dataset we found no adjustment in precision for that dataset i.e. the information in a standard dataset is now pre-handled and there is no need of further pre-processing. While by applying a similar component determination procedures on the dataset that is accumulated from the neighborhood doctor’s facility we found the sudden increment in the exactness of the model. At last, we reasoned that dependent on the precision the model created, the effective component determination calculation on the nearby dataset of the diabetes infection is the element choice method by utilizing Burota bundle. It successfully expands the exactness of the model with a superior.

References
[1] Hasan Akan, Ugur Demirak and Niyazi Kilic 2012 IEEE In Computers and Communications.
[2] Akhiljabbar M, Priti Chandra B L and Deekshatulu 2012 12th International Conference on Intelligent Systems Design and Applications (ISDA, IEEE).
[3] Sadhana Tiwari and Birmohan Singh 2015 Third International Conference on Image Information Processing, IEEE.
[4] Abu Bakr Soliman, Mohamed Fares, Mohamed M. Elhefnawi and Mahmoud Al-Hefnawy 2016 5th International Conference on Computers and Communications, IEEE.
[5] Suganya T S and S Murugavalli 2017 International Conference on Algorithm, Methodology models and Applications in Emerging Technologies, IEEE.
[6] Bin Wei, Qinke Peng, Xuejiao Kang and Chenyao Li 2010 Proceedings of the 8th World Congress on Intelligent Control and Automation, IEEE.
[7] Muthukumaran V and Ezhilmaran D 2020 International Journal of Information Technology and Web Engineering 15(4) pp 18-36.
[8] Nichole Challita, Mohamad Khalil and Pierre Beauséjour 2016 IEEE International Multidisciplinary Conference on Engineering Technology (IMCET), IEEE.
[9] Mohammed Siyad B, Visakh R and Abdul Nazeer K A 2017 International Conference on Innovations in Power and Advanced Computing Technologies.
[10] Ganesh Gopal Deverajan, V Muthukumaran, Ching-Hsien Hsu, Marimuthu Karuppihah, Yeh-Ching Chung and Ying-Huei Chen 2021 Transactions on Emerging Telecommunications Technologies.