Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 11579 - 11584
Received 05 March 2021; Accepted 01 April 2021.

Earlier Detection of Diabetes using Machine Learning

Ponnila P[1], Janani T[2], Deepika R[3]
[1, 2, 3] Bannari Amman Institute of Technology, Erode – 638 401, India
swathikapons@gmail.com

ABSTRACT

In worldwide group of people experience the illness and suffer due to the excess blood sugar in diabetes cause complications. It can severely damage to vital parts of the body like kidney, eyes and cardio vascular and it may have the risk of heart attack and stroke. Early detection of Diabetes plays a very important key role in healthcare. Our system is to efficiently identify the presence of blood sugar level with checking the number of parameters and by applying the machine learning techniques. Diabetes identification is mainly based on applying the classification algorithm may includes Logistic Regression, Naive Bayes, SVM, Decision Tree, Random Forest with by use feature selection algorithm which is used to solve the feature selection problem. The feature selection algorithm which is accomplished of removing the unwanted variable which is having low variance and also it is used to reduce the irrelevant and redundant features [1]. The tentative outcome show that the comparison of classifier algorithm with Random Forest, Logistic regression, SVM, Naive Bayes, , Decision tree and prove that the Logistic regression gives the good accuracy level and further the algorithm applied with the feature selection dataset which refers removal of low variance data in the dataset by finding the correlation between the variable in the dataset which gives the good accuracy level compared with the dataset which is not the process of selecting a subset of relevant features for use in model construction it means we are considering the all field in the dataset

Keywords: Diabetes identification, classification, features selection, disease diagnosis, intelligent system, medical data analytics

I. Introduction

The research study has the following contributions.

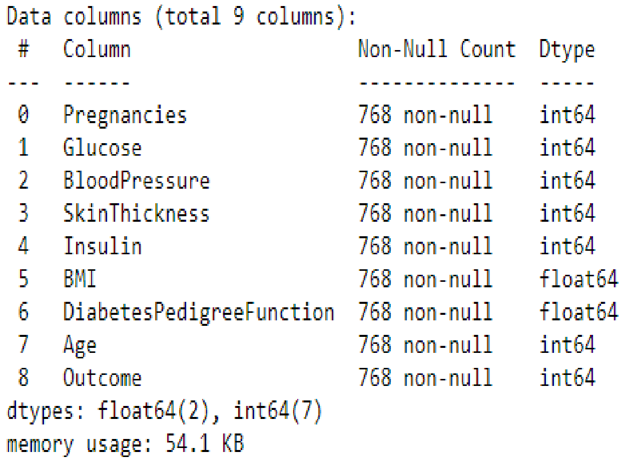
- Initially we are try to deal with the consequence of features selection by using pre-processing techniques and other features selection algorithms such as correlation which is used to find the relation between the data and the data which is having the low variance data may be dropped from the dataset which is not affect out result and to improve the accuracy level of the algorithm and testing of the classifiers algorithm with Random Forest, Logistic regression, SVM, Naive Bayes and Decision tree that identify which feature selection algorithm and classifier the info which gives good leads to term of accuracy and computation time.
- Secondly, the author is try to find the accuracy of the data by using the classifier algorithm with Random Forest, Logistic regression, SVM, Naive Bayes and Decision tree and prove that Logistic Regression will gave the good accuracy level without applying the subset of data (i.e) feature selection
- Thirdly, by using the feature selection algorithm that is Pearson correlation is used to find the weak features from the diabetes database which may degrade the accuracy of the classifiers.
- Last step, compare the result with the dataset by using classifier algorithm comparison and also the dataset which applied for feature selection that removes the low variance data which may affect the performance of the data.

- Shows the comparison result for the classifier algorithm with Logistic regression, SVM, NB, Random Forest and Decision tree with and without using the feature selection process

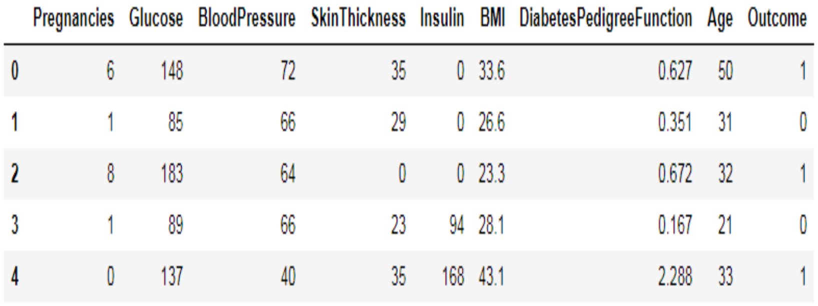
The paper remaining part is followed as. In section 2 the theoretical and mathematical formula of feature selection process and classification algorithms are discussed in details. In section 3 results of all experiments are analyzed and discussed in details. The last section 4 the conclusion and future direction of the research work are explored in details

II. DATASET DESCRIPTION

Dataset of Diabetes contains 9 columns which is described below:



```
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Pregnancies               768 non-null    int64
 1   Glucose                   768 non-null    int64
 2   BloodPressure             768 non-null    int64
 3   SkinThickness             768 non-null    int64
 4   Insulin                   768 non-null    int64
 5   BMI                       768 non-null    float64
 6   DiabetesPedigreeFunction  768 non-null    float64
 7   Age                       768 non-null    int64
 8   Outcome                   768 non-null    int64
dtypes: float64(2), int64(7)
memory usage: 54.1 KB
```



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

The aim of the work is to predicting the probability of the onset of diabetes using ML techniques. The accuracy of Random Forest, Logistic regression, SVM, Naive Bayes and Decision tree is compared which is categorized based on the supervised and unsupervised learning.

III. FEATURE SELECTION METHODS

Feature selection method is mainly used to reduce the number of input variable to those that are supposed to be most important or useful to a model to predict the outcome variable

We use the Indian dataset to predict the model of Diabetes Prediction using Machine learning algorithm and it is based on FS method. FS plays a very important role to find the important data which is very close to the target variable and this have to consider some of the predictive modelling problems have a huge number of input variables which can affect the performance of training models by occupying the large amount of system memory. And also it can degrade the performance when including input variables which is having the low variance with target variable.

Diabetes dataset are supervised data due to the use of target variable to remove the irrelevant data by applying the features selection methods which may divide into wrapper method and filter based method. The both selection model are supervised model which evaluate based the performance of resulting model.

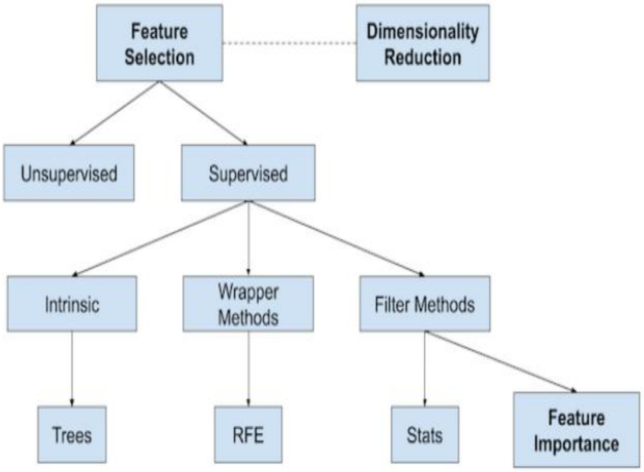


Figure 1: Features selection Techniques

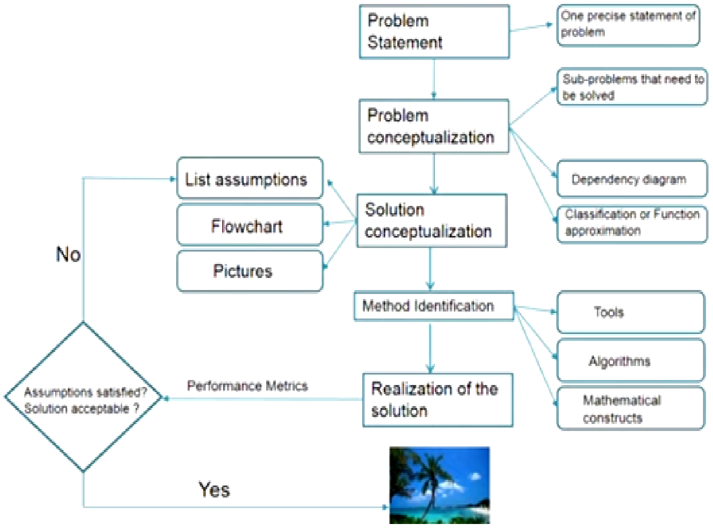

Problem Statement
One precise statement of problem
Problem conceptualization
Sub-problems that need to be solved
Dependency diagram
Classification or Function approximation
List assumptions
Flowchart
Pictures
Solution conceptualization
No
Method Identification
Tools
Algorithms
Mathematical constructs
Assumptions satisfied? Solution acceptable ?
Performance Metrics
Realization of the solution
Yes

Figure 2: Data analytics Framework

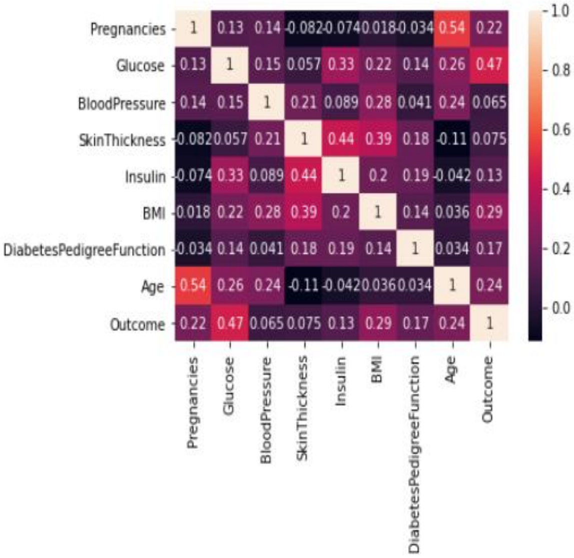
The correlation is comes under the filter method and it consider subset of data which is having highly correlated with the target data. The system is built after omitting the features i.e irrelevant data by applying the correlation that is pearson correlation. The filtering method is achieved by using finding dependency matrix and it is most using by the Pearson correlation

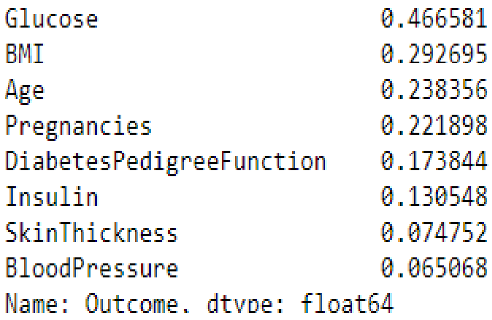
Received 05 March 2021; Accepted 01 April 2021.

The Pearson coefficient of correlation has lies between -1 to 1

- a worth closer to 0- weaker correlation (exact 0 implying no correlation)
- a worth closer to 1 - implies stronger direct correlation
- a worth closer to -1 - implies stronger indirect correlation

The correlation of the data with the target value of dataset is described below:

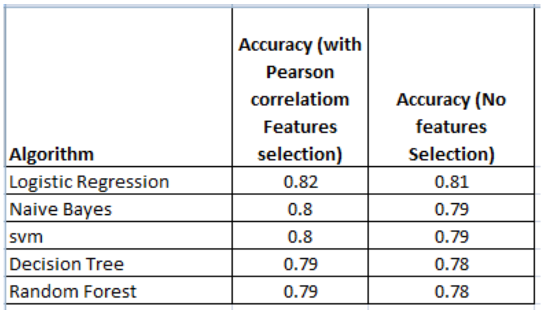




```
Glucose                     0.466581
BMI                         0.292695
Age                         0.238356
Pregnancies                 0.221898
DiabetesPedigreeFunction    0.173844
Insulin                     0.130548
SkinThickness               0.074752
BloodPressure               0.065068
Name: Outcome, dtype: float64
```

IV. RESULT AND ANALYSIS

The following are the algorithm comparison between different classifiers algorithm with Logistic regression, SVM, Naive Bayes, Random Forest and Decision tree with and without variable selection which means method of selecting a subset of relevant features



Algorithm	Accuracy (with Pearson correlatiom Features selection)	Accuracy (No features Selection)
Logistic Regression	0.82	0.81
Naive Bayes	0.8	0.79
svm	0.8	0.79
Decision Tree	0.79	0.78
Random Forest	0.79	0.78

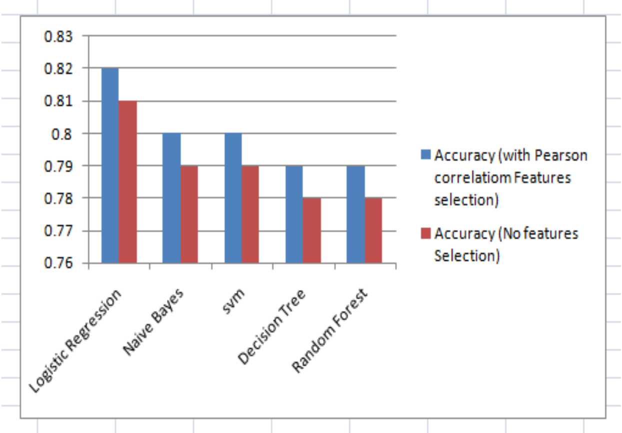


Figure 3: Comparison of Accuracy with different algorithm

V. Conclusion and Future Work

This paper gives the solution of predicting earlier diabetes prediction for the pregnancy women by applying classifier algorithm with Logistic regression, Support vector machine, Naive Bayes, Random Forest and Decision tree and prove that the Logistic regression gives the good accuracy level and further the algorithm applied with the feature selection dataset which refers removal of low variance data in the dataset by finding the correlation between the variable in the dataset which gives the good accuracy level compared with the dataset which is not the process of selecting a subset of relevant features for use in model construction it means we are considering the all field in the dataset. And our future work will be using different features selection process in order to improve the accuracy level of the result. The different features selection will be apply based on the numerical and categorical data will be chosen.

References

[1] Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare, jian ping li, amin ul haq , salah ud din , jalaluddin khan , asif khan and abdus saboor , VOL 8, June 9, 2020,IEEE Access

[2] Privacy Preservation using (L, D) Inference Model Based on Dependency Identification Information Gain, R. Deepika, V. Divya, C. Yamini, P. Sobiyaa, International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-8, Issue-6S3, September 2019

[3] An Efficient K-Means Clustering Initialization Using Optimization Algorithm V.Divya, R.Deepika, C. Yamini, P.Sobiyaa , 2019 International Conference on Advances in Computing and Communication Engineering (ICACCE), IEEE, 30 April 2020

[4] Brain Tumor Classification using Convolution Neural Network and Size Estimation by Marker Based Watershed Segmentation, Sathesh Kumar K., Arun Kumar R., Saranya S., Deepika R., Divya V. International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-6, March 2020

[5] M. Durairaj and N. Ramasamy, ‘‘A comparison of the perceptive approaches for preprocessing the data set for predicting fertility success rate,’’ Int. J. Control Theory Appl., vol. 9, no. 27, pp. 255–260, 2016.

[6] L. A. Allen, L. W. Stevenson, K. L. Grady, N. E. Goldstein, D. D. Matlock, R. M. Arnold, N. R. Cook, G. M. Felker, G. S. Francis, P. J. Hauptman, E. P. Havranek, H. M. Krumholz, D. Mancini, B. Riegel, and J. A. Spertus, ‘‘Decision making in advanced heart failure: A scientific statement from the American heart association,’’ Circulation, vol. 125, no. 15, pp. 1928–1952, 2012.

[7] S. Ghwanmeh, A. Mohammad, and A. Al-Ibrahim, ‘‘Innovative artificial neural networks-based decision support system for heart diseases diagnosis,’’ J. Intell. Learn. Syst. Appl., vol. 5, no. 3, 2013, Art. no. 35396.

[8] Q. K. Al-Shayea, ‘‘Artificial neural networks in medical diagnosis,’’ Int. J. Comput. Sci. Issues, vol. 8, no. 2, pp. 150–154, 2011.

[9] J. Lopez-Sendon, ‘‘The heart failure epidemic,’’ Medicographia, vol. 33, no. 4, pp. 363–369, 2011.