(IJRMST) 2016, Vol. No. 1, Jan-Jun **e-ISSN: 2455-5134, p-ISSN: 2455-9059**

DEVELOPING AN INTEGRATED ARTIFICIAL INTELLIGENCE MODEL BASED ON DATAMINING IN THE EARLY DETECTION, DIAGNOSIS AND PREDICTION OF DIABETES

Shubham Bhardwaj

National Institute of Technology, Hamirpur, Himachal Pradesh

ABSTRACT:

The process of selecting, categorizing, and evaluating transformations are all components called database knowledge discovery model, used to extract useful models and information from data. The two main types of machine learning algorithms are supervised and unsupervised. Unsupervised algorithms can draw interferences from data sets, whereas supervised learning algorithms use the experience to predict new or invisible data. Directed learning is likewise depicted as order. This study employs classification methods to produce a more precise classification.

The classification algorithm has been applied to the Indian Diabetes Dataset of the PIMA of the Public Foundation of Diabetes, Stomach-related and Kidney Sicknesses which contains information on diabetic ladies.

INTRODUCTION

Diabetes is prevalent and poorly managed, resulting in a high rate of preterm death. Maintaining a healthy blood sugar level can have significant health benefits and lower the risk of developing diabetes. Progressively, ceaseless observing of blood glucose is the actual test. Diabetes prediction monitoring immediately generates an emergency alert for preventative measures. However, medication may be misinterpreted if glucose levels are only monitored without considering other factors like an electrocardiogram and physical activity. We propose an artificial intelligence-based healthcare system for maintaining blood glucose levels to address the issues above. The experimental results demonstrate the experimental results demonstrate the improved performance of the proposed system in terms of energy efficiency, forecasting accuracy, computational complexity, and latency.

IMPLEMENTATION

A. Details About the Dataset This project uses a PIMA dataset downloaded from Kaggle.com.

This dataset has 9 credits and a sum of 768 records.

From this dataset, we can learn about a person's age, glucose level, number of pregnancies, sugar level, blood pressure level, skin thickness, and body mass index (BMI).

(IJRMST) 2016, Vol. No. 1, Jan-Jun **e-ISSN: 2455-5134, p-ISSN: 2455-9059**

B. Data Pre-processing For improved dataset analysis, it is critical to pre-process our data. During this procedure, in this procedure, we checked several rows and columns in our dataset. Additionally, we checked a number of our dataset's null values.

C. Data Cleaning In data cleaning, all of the null values in our dataset are filled in. A few columns in our dataset, including glucose, blood pressure, skin thickness, BMI, and insulin, have null values. As a result, the attribute's median value takes the place of all the missing values.

D. Feature Selection Feature selection is a method in which features highly correlated with one another are removed. Dropping constant features, correlation, and other methods are among the many methods used in feature selection. The correlation was used in this case. We imported seaborn and performed the entire correlation step with the assistance of a heat map.

E. Algorithms Because this problem statement is related to supervised learning, we tested our model with a variety of classifiers to determine which algorithm produces the lowest error. Logistic Regression, Xgboost, and a Random forest classifier were utilized here here.Finally, we discovered that Logistic Regression produced a lower error rate for our model after employing all three algorithms.

F. Cross-Validation The two processes that makeup cross-validation are cross_val_score and cross_val_predict. Here, the Cross_val_score function performs cross-validation. It requires cv=3, which indicates that it will train our model on the first two datasets and predict the value on the third; Our model will then be tested on the sixth dataset and trained on the fourth and fifth datasets, respectively. We also have scoring = "accuracy" in this case, indicating that accuracy metrics should be used for scoring. The cross_val_predict function now reveals that we are obtaining this kind of prediction we are obtaining this kind of prediction by training the dataset in this manner.

G. Confusion Matrix This matrix requires both our predicted and training data. Our data are labelled Y_train. It requires both our actual prediction values and them. Several of the positive and negative predictions are accurate, while others are incorrect. This will be a case of the perfect confusion matrix if we have predicted all correct values.

H. Precision and Recall It will also consider our actual prediction and the values we predicted. The precision-recall function was then utilized. This demonstrates that increasing precision decreases recall, whereas decreasing precision increases recall. In this case, we also have a term called a threshold, which means that if the value is higher than this value, it will be positive, and if it is lower, it will be negative. On the x-axis, precision and recall are plotted against the threshold in this instance. Y_scores will provide the decision threshold for logistic regression.Cross_val_predict's parameters are identical to ours here. We don't want accuracy, and that's all that has changed; Here, we need a decision function.Y_scoresmeans we are getting the limits.

I. F1 Score The harmonic mean of recall and precision is the F1 score. Recall decreases as precision increases and recall increases as precision decreases. Precision and recall are sacrificed here. Therefore, for better analysis, we have also calculated the F1 score.

(IJRMST) 2016, Vol. No. 1, Jan-Jun **e-ISSN: 2455-5134, p-ISSN: 2455-9059**

The precision-recall curve is being plotted in this location.Plt. the plot gives the edge, accuracy and b- - to recognize the accuracy curve and name = accuracy. We gave the location as upper left using the legend function. Since we want to stay between 0 and 1, we used ylim here. Additionally, we have used [:1], indicating that we wish to eliminate the final value from precision and recall.

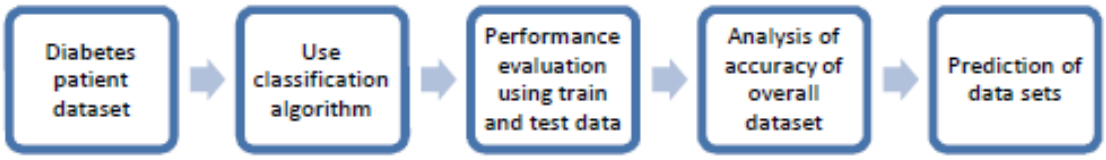
Precision is the ratio of the corrected and made total positive predictions to the total positive predictions. The recall is the ratio of our total positive observations to the total positive observations that our classifier found to be correct.

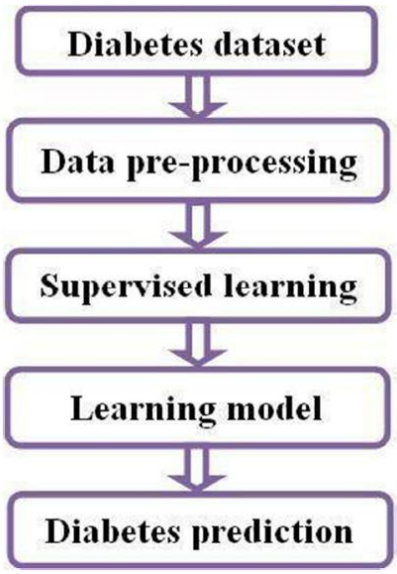
J. Algorithms To better analyse our model, we used the logistic regression classifier, the random forest classifier, and the xgboost classifier here.

Last but not least, the logistic regression classifier produced fewer errors. Logistics predicts the probability.

We used those probabilities to classify. This is a class of generalized linear model algorithms.

FLOWCHART





CONCLUSIONS

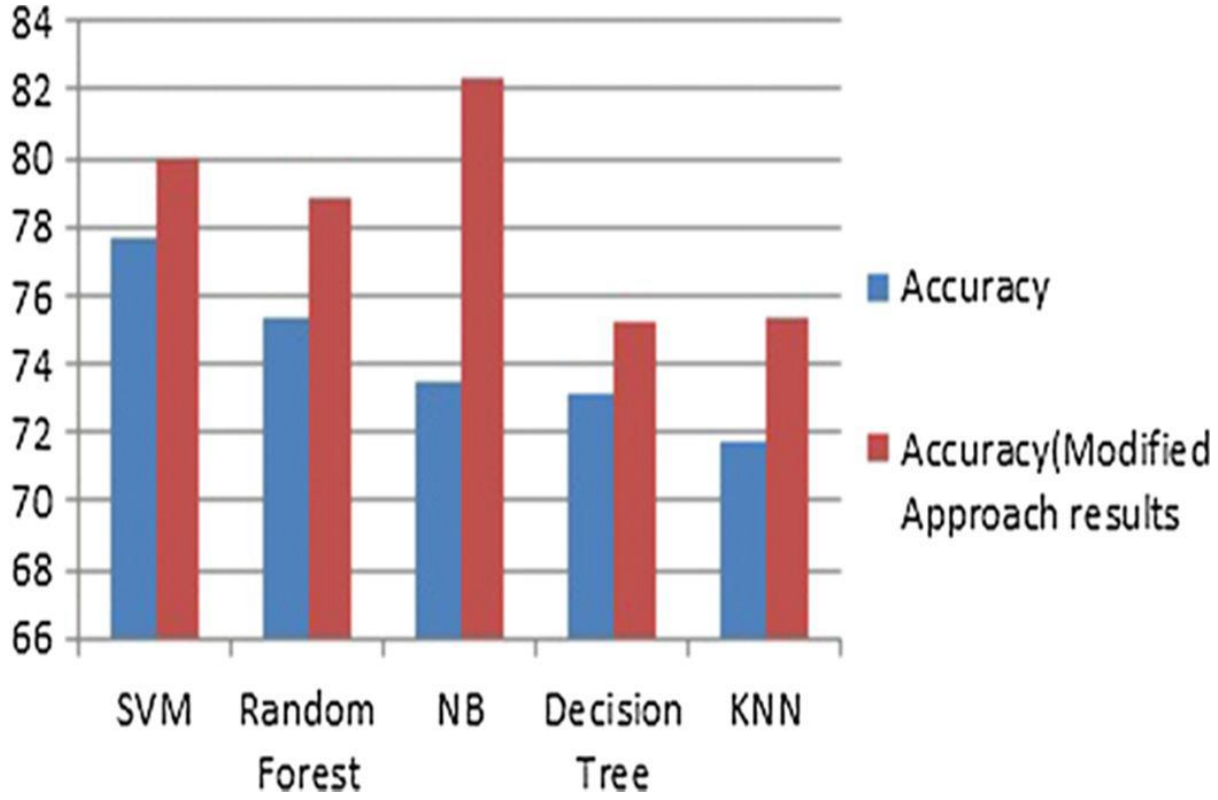
The development of an ideal and effective structure for diabetes prediction is the primary objective of our work.

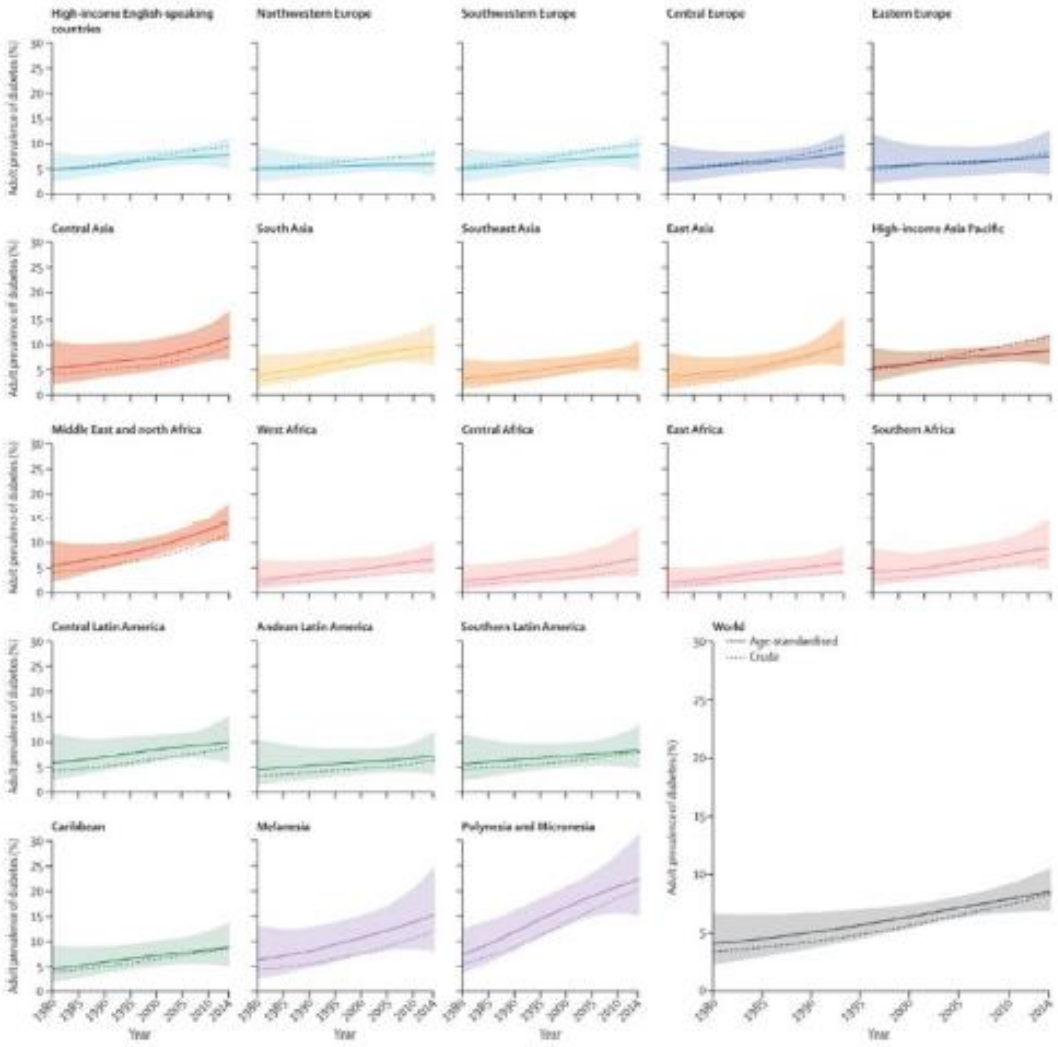
Ours PCA for dimensionality reduction, k-means for clustering, and logistic regression for classification after carefully analysing other published and ongoing research.

We first applied the PCA method to our dataset to improve other people's k-means results. Even though PCA is a well-known method that is also effective at monitoring k-means clustering, the logistic regression classification model requires more attention.

However, our experiment has demonstrated that incorporating PCA and k-means may be effective in a well-managed logistic regression model for diabetes prediction. The knowledge of the study includes getting a better k-means cluster result than other researchers have concluded from similar studies.

When predicting diabetes onset, the logistic regression model performed even better than other algorithms when applied to our phenomena and other onsets.






```
# 76% accuracy.
a.mean()
```

```
Out[72]:

0.7432432432432433
```

REFERENCES

[1] Retrieved http://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-ofdeath, Accessed date: 27 July 2018.

[2] http://www.who.int/news-room/fact-sheets/detail/diabetes retrieved 27/07/ 2018.

[3] https://www.diabetesdaily.com/learn-about-diabetes/what-is-diabetes/how-manypeople-have-diabete

[4] Tarun Jhaldiyal, Pawan Kumar Mishra Analysis and prediction of diabetes mellitus using PCA, REP and SVM 2014 Int J Eng Tech Res (IJETR) ISSN: 2321-0869, Volume-2, Issue-8.

[5] Prabhu P, et al. Improving the performance of K-means clustering for high dimensional data set. Int J Comput Sci Eng June 2011;3

[6] ISSN: 0975-3397. [6] Khandegar Anjali. Khushbu Pawar diagnosis of diabetes mellitus using PCA, neural Network and cultural algorithm. Int J Digital Appl Contemp Res 2017;5(6).

[7] Novakovic J, Rankov S. Classification performance using principal component analysis and different value of the ratio R. Int J Comput Commun Control 2011;Vol. VI(2):317–27. ISSN 1841-9836, E-ISSN 1841-9844.

[8] Motka Rakesh, Parmarl Viral, Kumar Balbindra, Verma AR. Diabetes mellitus forecast using different data mining techniques. IEEE 4th international conference on computer and communication technology (ICCCT). IEEE; 2013. p. 99–103.

[9] https://en.wikipedia.org/wiki/K-means_Clustering.

[10] Seyed S, Mohammad G, Kamran S. Combination of feature selection and optimized fuzzy apriori rules: the case of credit scoring. Int Arab J Inf Technol2015;12(2).