iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** IX **Month of publication:** September 2022

DOI: https://doi.org/10.22214/ijraset.2022.46633

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue IX Sep 2022- Available at www.ijraset.com

Diabetes Prediction System

Ishaan Devendra Kalbhor
Student, Department of Computer Science & Engineering, MIT-ADT University, Pune, India

Abstract: This model of retrieving useful information and models from the data is also called database knowledge discovery, which involves certain phases such as data selection, classification and transformation evaluation. Machine learning algorithms are primarily classified as supervised and unsupervised. A supervised learning algorithm uses past experience to make predictions about new or invisible data, whereas unsupervised algorithms can draw interferences from data sets. Supervised learning is also described as classification. This study uses classification technique to produce a more accurate belong to class. The classification algorithms have been applied to the Indian Diabetes Dataset of the PIMA of the National Institute of Diabetes, Digestive and Kidney Disease which contains data on diabetic women.
Keywords: AI, Deep Learning, Computer Vision

I. INTRODUCTION

Diabetes has a high prevalence and low control, leading to a high rate of premature mortality. Maintenance of blood sugar can provide significant health benefits and reduce the risk of diabetes.
In real time, continuous monitoring of blood glucose is the primary challenge.
However, monitoring glucose levels alone without consideration of other factors such as ECG and physical activities may mislead medication. By implementing diabetes prediction monitoring, an emergency alert is generated immediately for precautionary actions. The experimental results show the improved performance of the proposed system in terms of energy efficiency, forecasting accuracy, computational complexity and latency. To solve the above problems, we are proposing an energy-efficient artificial intelligence health care system to maintain blood glucose**.**

II. IMPLEMENTATION

A. Dataset Details

The dataset used in this project is a PIMA dataset, taken from Kaggle.com. This dataset has 9 attributes and a total of 768 records. This dataset gives us the information of person's age, glucose level, number of pregnancies, sugar level, blood pressure level, skin thickness and BMI.

B. Data Preprocessing

In this process we have checked number of rows and number of columns in our dataset. We have also checked number of null values in our dataset. It is very important to preprocess our data for better analysis of our dataset.

C. Data Cleaning

Data cleaning basically means filling all the null values present in our dataset. In our dataset few column has null values like dataset, glucose, blood pressure, skin thickness, BMI, insulin. So all the missing values are being replaced with the median value of that attribute.

D. Feature Selection

Feature selection is technique where we remove those features which are highly correlated with each other. Feature selection involves various techniques like dropping constant feature, correlation, etc. Here we have used correlation. We have imported seaborn and with the help of heat map, we did the whole correlation part.

E. Algorithms

As this problem statement lies inside supervised learning, so we have used various classifier to test our model, and to check on which algorithm our model gives less error. We have used here Logistic Regression, Xgboost, Random forest classifier. And finally after applying all three algorithm we found that our model gives less error with Logistic Regression.

F. Cross Validation

Cross validation has two process in it, cross_val_score and cross_val_predict. Cross_val_score function does cross validation over here. It takes cv=3, which means that on first two dataset it will train our model and on the third one it will predict the value, then again on fourth and fifth dataset our model will be trained and on sixth it will be tested. Here we also have scoring = "accuracy", which means we want accuracy metrics as scoring. Now cross_val_predict function tells us that, by training the dataset in this way we are getting this type of predictions.

G. Confusion Matrix

This matrix wants our training data and our predicted data. Y_train is the label of our data. It takes the real prediction and our prediction values. Basically it gives us number of correct negative and positive prediction and number of incorrect negative and positive predictions. If in case, we have predicted all correct values, then this will be a situation of perfect confusion matrix.

H. Precision and Recall

It will also take our real prediction and our prediction values. Then we used precision recall function. Through this we get to know that if we increase precision then recall gets decreased and if we decrease precision then recall gets increased. Here we also have a term called threshold, which means that if the value is greater than this value then it will be positive, and if the value is less than this value then it will be negative. Here precision and recall are being plotted vs threshold on x axis. Y_scores will give the decision threshold that logistic regression is using. Here we have same parameters as cross_val_predict. Only change is that here we do not want accuracy, we want decision function. Y_scoresmeans, we are getting the thresholds.

I. F1 Score

F1 score is the harmonic mean of precision and recall. When we increase precision then recall gets decreased, and when we decreased precision then recall gets increased. This is precision recall trade off. So we have also calculated F1 score for better analysis.

Precision_recall_curve

Here we are plotting the precision recall curve. Plt.plot

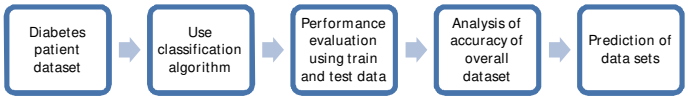
gives the threshold, precision and b--, to identify precision

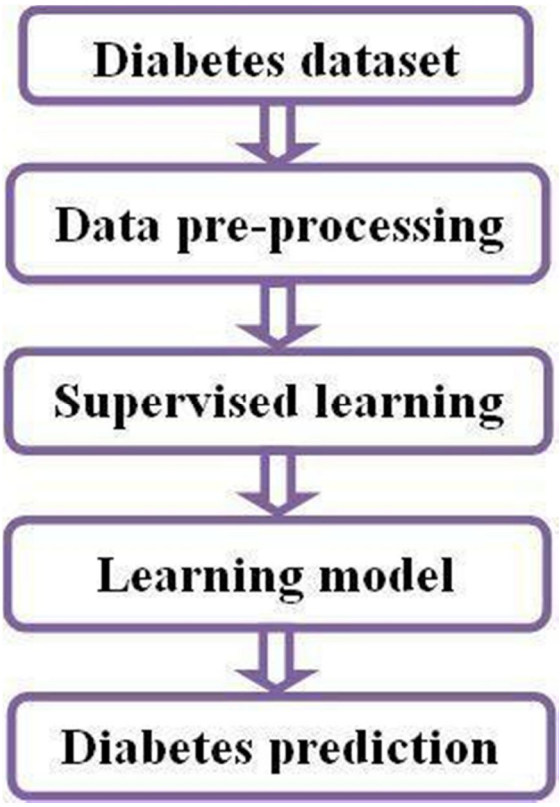
curve and label = precision. Through legend function we have given location as upper left. We have used here ylim because we want to stay in between 0 and 1. We have also used [:1], which means we want to remove last value from both precision and recall. Precision is the ration of total positive predictions which are corrected and total positive predictions we made. Recall is the ratio of total positive observations which our classifier found correct and total positive observations we made.

J. Algorithms

Here we have used logistic regression classifier, random Forest classifier and xgboost classifer for better analysis of our model. Finally we got less amount of error from logistic regression classifier. Logistic predicts the probability. With those probabilities we did classification.This is an algorithm of generalised linear model class.

III. FLOWCHART





IV. CONCLUSIONS

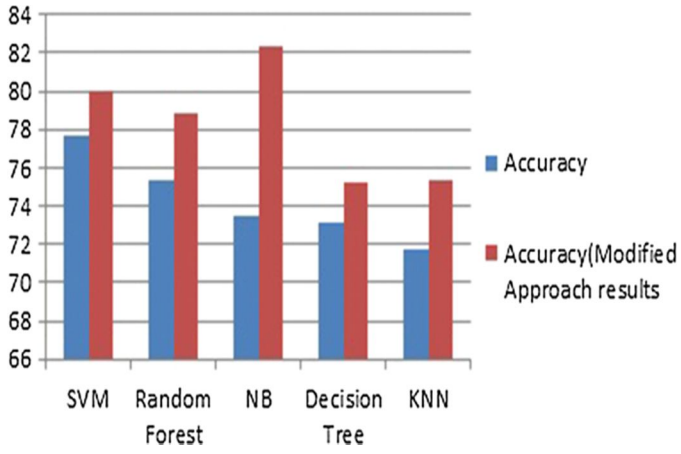
The main aim of our work is to design a perfect efficient structure for the diabetes prediction.

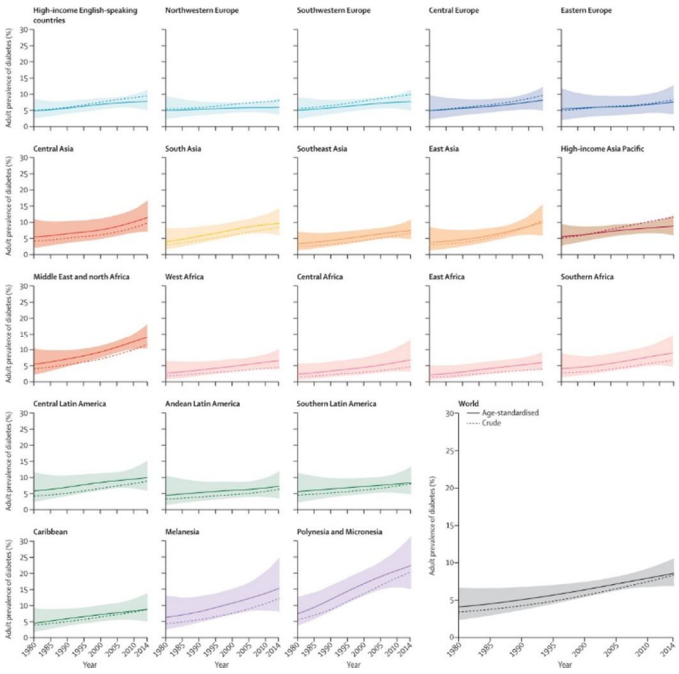
After doing careful understanding of other on going and available work, we come to a conclusion that , our model which consists of using PCA for dimensionality reduction, k-means for clustering, and logistic regression for classification.

With the intention to improve the k-means result of other people, we firstly applied the PCA technique to our dataset although PCA is a well-known way, also this is efficient in monitoring k-means clustering and in turn the logistic regression classification model has not been given sufficient response.

however the experiment that we have just shown that a well managed logistic regression model for predicting diabetes is possibly working with the integration of PCA and k-means. The knowledge achieved in the study includes, the ability to obtain an improved k-means cluster result above what other people have concluded in similar studies.

Also logistic regression model worked at a well improved level via predicting diabetes onset, as compared to the values obtained when other algorithms were used in our phenomena and that of other onsets.






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

```
Out[72]:
0.7432432432432433
```

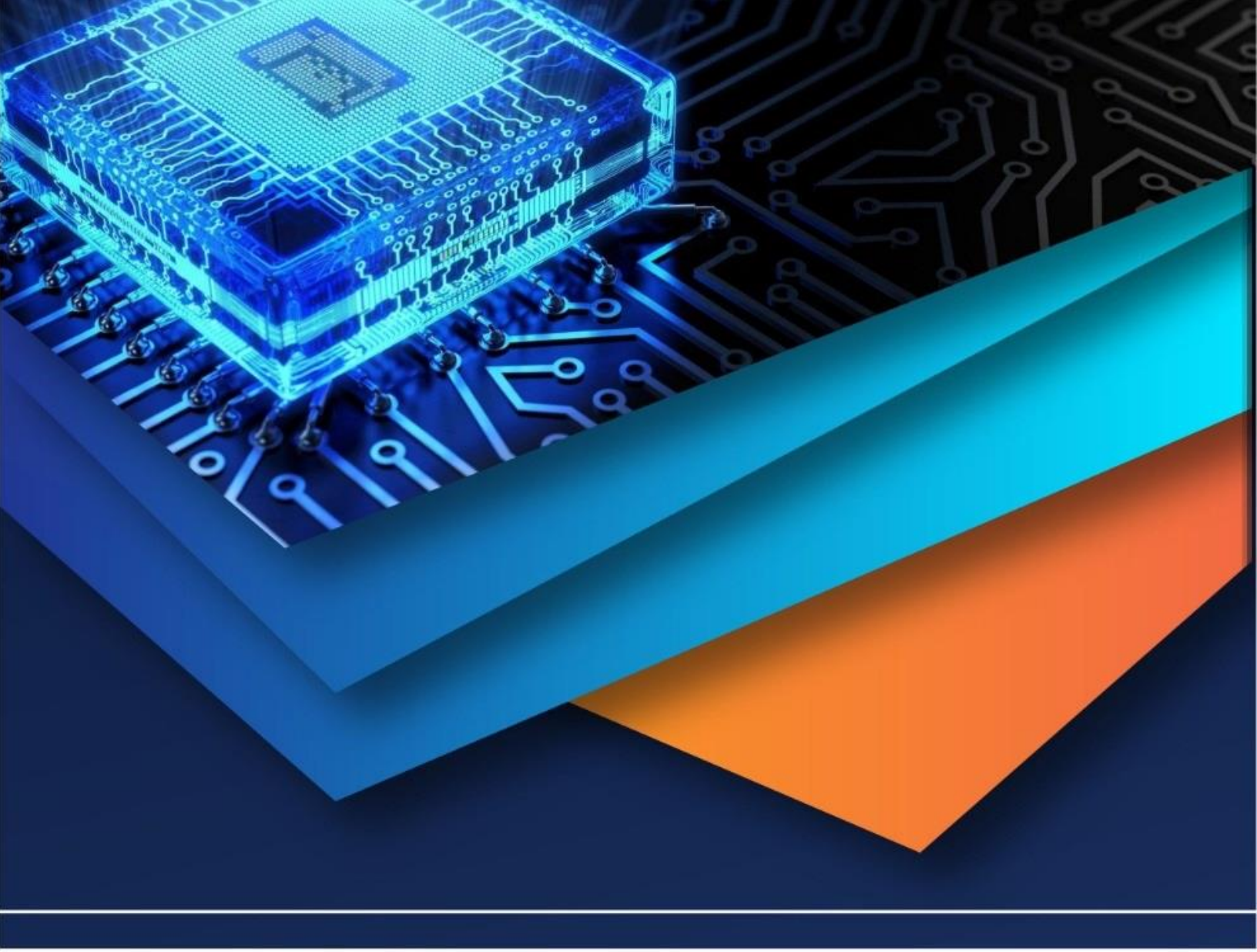
V. FUTURE SCOPE AND ENHANCEMENTS

Our first module Diabetes Prediction System is successfully implemented. We can still try to increase its accuracy and can deploy for applications where it is useful/needed. In the second application Diabetes Prediction System, We gave input as image, so we can give it live input and in this application also we can try to increase accuracy. Also we can integrate warnings and alert in this application if people in frame are violating deadlines (i.e. not aware with the diabetic consequences).

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 Technol 2015;12(2).

****



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY