Improved Models for Diabetes Prediction by Integrating PCA Technique

Lu Yao*

Hwa Chong Institution, Singapore, Singapore

*Corresponding author: 182127U@student.hci.edu.sg

Abstract. Diabetes is an incurable, chronic disease indicated by hyperglycemia. It may lead to severe complications and is consequently one of the top ten causes of mortality in recent years. Fortunately, early detection can significantly aid in disease management. Based on their regular medical checkup results, people can make a preliminary evaluation of their risk of having the condition with the help of machine learning techniques. In this study, diabetes is predicted using Extreme Gradient Boosting (XGB), Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR), Decision Tree (DT), and K-Nearest Neighbor Algorithm (KNN), ranked in descending order of accuracy. XGB comes out on top with an accuracy of 87.29%. The dimensionality reduction approach Principal Component Analysis (PCA) is conducted, resulting in a maximum accuracy gain of 1.27%. The study's enhanced diabetes prediction models allow people to more accurately gauge their risk of developing the condition. More potential patients would be notified to get checked, enhancing the disease's early detection rate.

Keywords: machine learning, PCA, diabetes.

1. Introduction

1.1. Background

Diabetes is a chronic, incurable condition characterized by hyperglycemia. As reported by the Centers for Disease Control and Prevention (CDC), there are approximately 531 million diabetics worldwide. Based on its rising mobility in recent years, this figure is anticipated to reach 642 million by 2040, implying that one out of every 10 adults will suffer from this disease in the future. Diabetes can induce a variety of serious complications, including diabetic ketoacidosis, and thus accounts for one fatality every five seconds. To make the problem worse, not all symptoms appear in the early stages, leaving patients unaware of their illnesses. As a result, 46.5% of people with diabetes are undiagnosed.

All the data points to a worrying future for this illness. Fortunately, if proper medical treatment is received on time, people can fully recover from pre-diabetes (i.e., the precursor stage of diabetes). Machine learning, among other things, can assist people in making preliminary judgments and predictions concerning diabetes mellitus based on the results of their regular health examinations.

1.2. Related Research

Numerous models have been employed to predict diabetes [1-5]. Zou et al. [2] constructed models based on Decision Tree, Neural Network and Random Forest and tested them using the classic "Pima Indians Diabetes" dataset, which gave the highest accuracy of 80.84%. Pradhan et al. [3] applied Artificial Neural Network for prediction, which gave an accuracy of 85.09%. El-Jerjawi and Abu-Naser [4] used the same method and obtained an accuracy of 87.3%. Wu et al. [5] adopted eight machine-learning models, which gave the highest accuracy of 79.49%. Some experts also tried out some dimensionality reduction methods to improve the models. By combining K-means and PCA techniques with the logistic regression model, Zhu, et al. [6] enhanced the accuracy of prediction by 1.98%. Duygu and Esin [7] applied Linear Discriminant Analysis (LDA) to decrease the dimensionality and extract features before generating a model.

1.3. Objection

This study constructs several models (LR, SVM, KNN, DT, RF, XGB) based on the CDC dataset and evaluates each one's capacity in diabetes prediction to determine the most suitable one. It also aims to measure the efficacy of PCA in improving the quality of every model.

2. Methodology

2.1. Data

2.1.1 Source of Dataset

A dataset from Kaggle [8] is used for diabetes prediction. The data consists of 253,680 responses to the Centers for Disease Control and Prevention (CDC)'s 2015 Behavioral Risk Factor Surveillance System (BRFSS) Survey, with 21 out of 454 features (Chapter 3.1) selected for analysis based on the literature on diabetes risk factors [9]. The dependent variable was a binary feature asking whether the person has diabetes, with "1" representing "yes" and "0" representing "no".

2.1.2 Data Mining Toolbox

Anaconda is an open-source and free Python programming language toolbox. It includes more than 250 commonly used packages for machine learning applications and data science. Applying these packages, the proposed model can be built on the dataset.

2.1.3 Data Preparing

Firstly, missing values and duplicated data are eliminated.

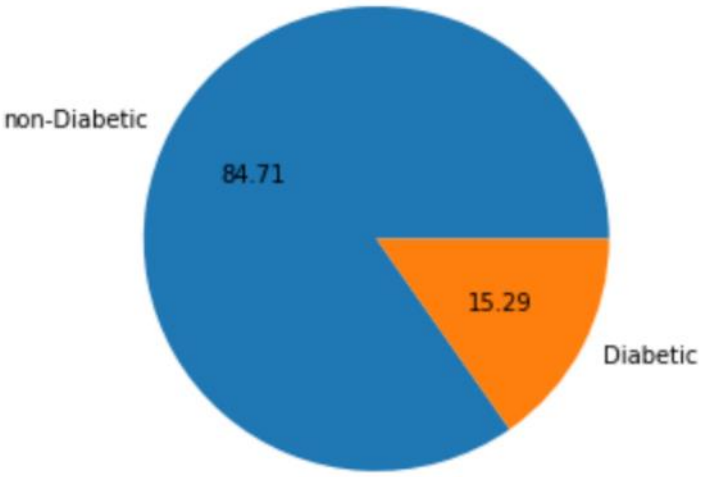


Fig. 1 Percentage of two classes (Photo credit: Original)

As observed in Fig. 1, the dataset contains many more non-diabetics than diabetes patients. To avoid biased prediction due to the normal class's dominance, the amounts of samples from each class in the dataset are then equalized before modeling. This leaves 35097 samples in both classes.

Finally, all the variables are standardized to avoid biased results caused by discrepancies between the initial ranges of variables. The scaled standardization is done by Formula (1):

$$Z = \frac{value - mean}{standard\ derivation} \tag{1}$$

2.2. Models

2.2.1 Logistic Regression

Based on the book Applied Logistic Regression [10], logistic regression with multiple predictors has the probability function p(X) of Formula (2):

$$p(x) = \frac{e^{\beta_0+\beta_1 X_1+\ldots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\ldots+\beta_p X_p}} \tag{2}$$

With X_i representing the p features and β_i representing the parameters. After the logit transformation, we derived Formula (3), which is linear in its parameters:

$$g(x) = \log\left(\frac{p(X)}{1-p(X)}\right) \tag{3}$$

$$= \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$$

The maximum likelihood approach is used to compute estimators for all parameters to build the best-fitted model. In other words, estimates like $\hat{\beta}_0$ are chosen to maximize the likelihood function (Formula (4)):

$$l\left(\hat{\beta}_0, \hat{\beta}_1 \ldots \hat{\beta}_p\right) = \prod_{i:\, y_i=1} p(x_i) \prod_{i\prime:\, y_i=1} \left(1-p(x_{i\prime})\right) \tag{4}$$

A person will be classified as a diabetes patient if the estimated probability is greater than 0.5.

2.2.2 Support Vector Machine

Support Vector Machine is a supervised classification model of machine learning.

Data in the dataset are first represented as $(X_i,\ Y_i)$, where X_i is a p-dimensional feature vector providing the details of the p features and Y_i is a label that, depending on its value, indicates the class that every point belongs to.

To separate the two classes, SVM wants to compute a "maximum-margin hyperplane" of dimension $(p-1)$, which is defined in a way that maximizes the distance between it and the nearest point from either class [11]. The hyperplane is in the form of Formula (5):

$$W^T X + b = 0 \tag{5}$$

where W is the weight vector, X is the input feature vector, and b is the bias.

The nearest points are called the support vectors, and the two hyperplanes containing these points are written as Formulas (6) and (7):

$$W^T X + b = 1 \tag{6}$$

$$W^T X + b = -1 \tag{7}$$

which are parallel to the hyperplane (5).

The distance between hyperplanes (6) and (7) is calculated to be $\frac{2}{\|W\|}$ which is maximized to obtain W. By substituting the W into hyperplane (4), the best hyperplane for the decision boundary is achieved.

2.2.3 K-Nearest Neighbor

K-Nearest Neighbor is a non-parametric supervised machine learning method specifically for classification [12].

The initial form of data is $(X_i,\ Y_i)$ as well. An unlabeled point is classified by assigning the label that appears most frequently among the k training samples nearest to it. The value of k is automatically determined and varies for different models to achieve the best classification accuracy. By trial, k is set to be 11 in this study.

2.2.4 Decision Tree

Another non-parametric supervised machine learning method is the decision tree [13].

It is a tree-like model in which each internal node represents a "test" on a characteristic (e.g., whether the individual has high blood pressure or not), each branch reflects the outcome, and each

leaf node shows the class label, which in this research is the diabetes prediction outcome. Classification rules are represented by the pathways from root to leaf.

A basic decision tree model for diabetes prediction employing three health markers is shown in Fig. 2.

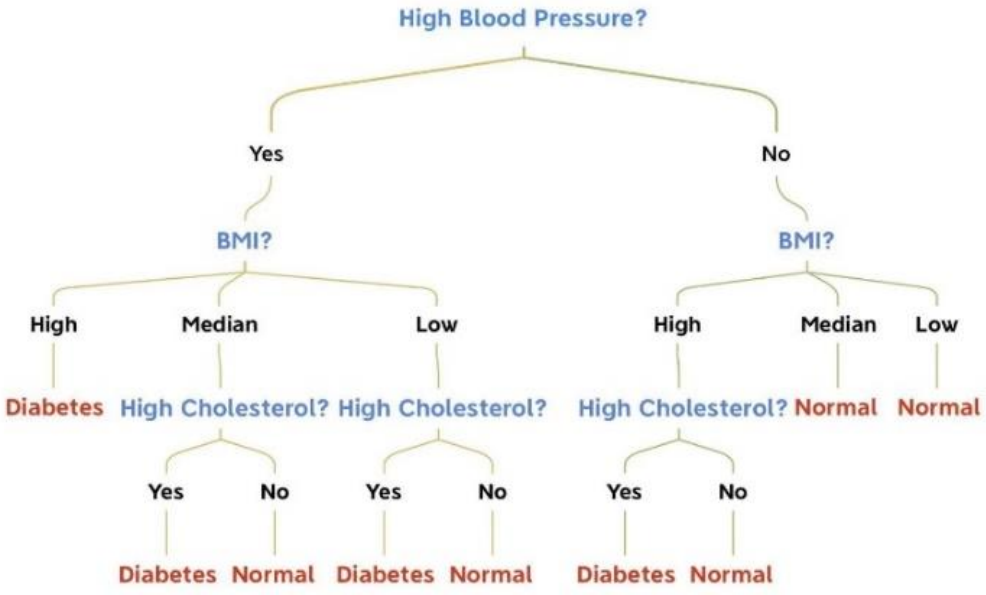


Fig. 2 Sample decision tree (Photo credit: Original)

Pruning is frequently done to avoid over-fitting caused by particular circumstances. In this study, the maximum depth of the tree is set at 12 to prevent the generation of non-significant branches.

2.2.5 Random Forest

Random Forest, which is based on decision trees and the bagging algorithm, was proposed by Breiman in 2001 [14].

"Bagging" refers to the process through which RT builds B decision trees ($\hat{f}^b(x)$), each of which "grows" depending on different training samples and features. Thus, every random tree will offer its own classification outcome when predicting the object based on some features. Then, the final output of RF will be the categorization that occurs most frequently among these results. For regression, the result will be given by Formula (8):

$$\hat{f}_{bagging}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^b(x) \tag{8}$$

2.2.6 Extreme Gradient Boosting

Extreme Gradient Boosting is a widely used scalable machine-learning algorithm for tree boosting [15].

Gradient boosting is the technique of combining a variety of weak decision tree models to enhance their collective accuracy. It sequentially trains a group of shallow trees $\hat{f}^b(x)$), with each iteration using the error residuals from the previous model to fit the next model. The final result is the weighted average of all of the predictions from the weak models (Formula (9)):

$$\hat{f}(x) = \frac{1}{B}\sum_{b=1}^{B}\lambda\hat{f}^b(x) \tag{9}$$

2.3. Dimensionality Reduction

Principal Components Analysis is a dimensionality reduction method that can be broken down into four steps.

To begin, the covariance matrix is constructed in order to comprehend the correlations between each pair of variables. Covariance and a sample 3x3 covariance matrix are calculated by Formulas (10) and (11) respectively:

$$Cov(X,Y) = \sum_{i=1}^{N} \frac{(x_i - \overline{x}) - (y_i - \overline{y})}{N} \tag{10}$$

$$\begin{bmatrix} Cov(X,X) & Cov(X,Y) & Cov(X,Z) \\ Cov(Y,X) & Cov(Y,Y) & Cov(Y,Z) \\ Cov(Z,X) & Cov(Z,Y) & Cov(Z,Z) \end{bmatrix} \tag{11}$$

Secondly, p pairs of eigenvalues and eigenvectors of the covariance matrix are computed using Formula (12):

$$A_i X = \lambda_i X \tag{12}$$

Where A_i is the eigenvector and λ_i is the corresponding eigenvalue.

Following that, we select q eigenvectors with the greatest matching eigenvalues to be the columns of the feature vectors and calculate the final dataset using Formula (13). The principal components are the columns of this dataset.

$$FinalDataSet = FeatureVector^T \times StandardizedOriginalDataSet^T \tag{13}$$

Finally, we plot prediction models using the final dataset.

2.4. Model Quality Measurement

The quality of a model is measured by four values, whose derivation formulas are given in Table 1.

Table 1. Quality Measurement

Measures	Description	Formulas
Accuracy(A)	Percentage of correctly predicted instances	(TP+TN)/Total number of sample
Precision(P)	Percentage of real patients in the diagnosed people (Correctness)	TP/(TP+FP)
Recall(R)	Percentage of patients being successfully diagnosed (sensitivity/Completeness)	TP/(TP+FN)
F-Measure (F1)	Percentage of patients being successfully diagnosed (sensitivity/Completeness)	2*(P*R)/(P+R)

FP: false positive TP: True positive FN: False negative

All four values are between 0 and 1, with a greater number indicating higher model quality.

3. Results

3.1. Data

Table 2 contains the detailed descriptions and data codes of all 21 features selected and the outcome variable.

Table 2. Details of Features

Features	Description	Type
Diabetes binary	Do you have diabetes?	0: No, 1: Yes
HighBP	Do you have high blood pressure?	0: No, 1: Yes
High Chol	Do you have high cholesterol?	0: No, 1: Yes
Chol Check	Have you had your cholesterol checked in the last 5 years?	0: No, 1: Yes
BMI	What is your Body Mass Index?	Continuous
Smoker	Have you smoked at least 100 cigarettes until now?	0: No, 1: Yes
Stroke	Have you had a stroke?	0: No, 1: Yes
HeartDiseaseorAttack	Do you have a myocardial infarction or coronary heart disease?	0: No, 1: Yes
Phys Activity	Have you done any physical activity in the past 30 days?	0: No, 1: Yes
Fruits	Do you consume fruit at least once a day?	0: No, 1: Yes
Veggies	Do you consume veggies at least once a day?	0: No, 1: Yes
Hvy AlcoholConsump	Do you have more than 14 drinks per week? (7 for women)	0: No, 1: Yes
AnyHealthcare	Do you have health care coverage of any type (e.g., health insurance)?	0: No, 1: Yes
NoDoc bcCost	Have you needed to see a doctor in the last 12 months but were unable to do so due to cost?	0: No, 1: Yes
GenHlth	Rate your health in general?	1: excellent, 2: very good, 3: good, 4: fair, 5: poor
MentHlth	How many days of poor mental health have you had in the past 30 days?	Discrete
PhysHlth	How many days have you had a physical illness or injury in the past 30 days?	Discrete
DiffWalk	Do you have serious difficulty walking or climbing stairs?	0: No, 1: Yes
Sex	What's your biological sex?	0: Female, 1: Male
Age	Which age category are you in? (13 categories in total)	1: 18-24, 9: 60-64, 13: 80 or older, etc.
Education	Which education category are you in? (6 category in total)	1: Never attended school/ kindergarten, 2: elementary school, etc.
Income	Which income bracket are you in? (8 category in total)	1: less than $10,000, 5: less than $35,000, 8: $75,000 or more, etc.

Fig. 3 is a heat map showing the correlation between every pair of features and the outcome variable.

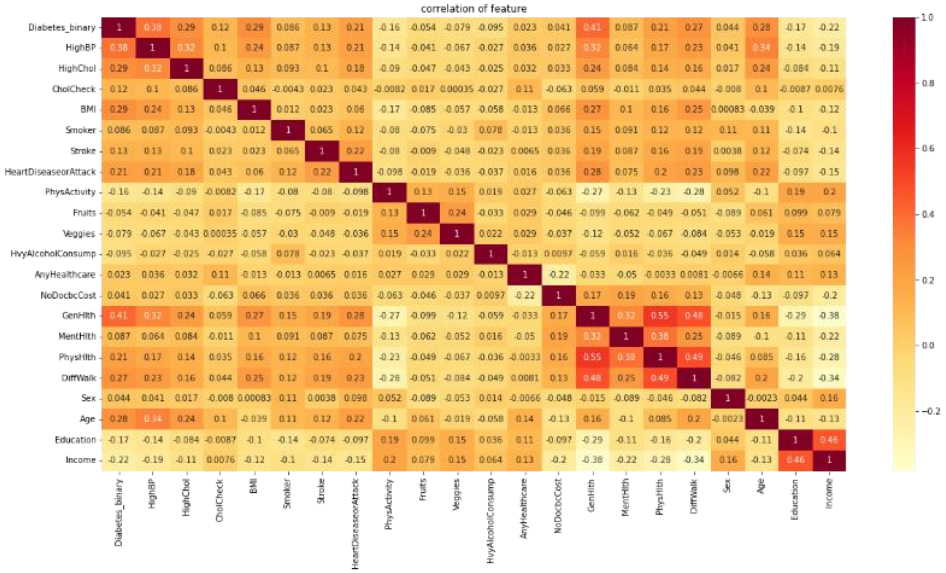


Fig. 3 Correlation of features (heat map) (Photo credit: Original)

General health, blood pressure, and BMI are the three most important indications of diabetes, according to their strong relationships with the disease.

3.2. PCA

Among the 21 features, several (e.g., healthcare coverage) show a correlation coefficient with diabetes of less than 0.05. This may limit the accuracy of models since the error of the test usually grows as the dimensionality of the model (i.e., the number of features or predictors) increases unless the variables added in are actually related to the result [16]. Additionally, some feature pairings are highly connected, with correlation coefficients greater than 0.5. Thus, principal component analysis (PCA) is conducted.

Table 3. Cumulative explained variance ratio

No. of features	1	2	…	17	18	…	21
Explained variance ratio	17.24	25.36	…	90.86	93.47	…	100

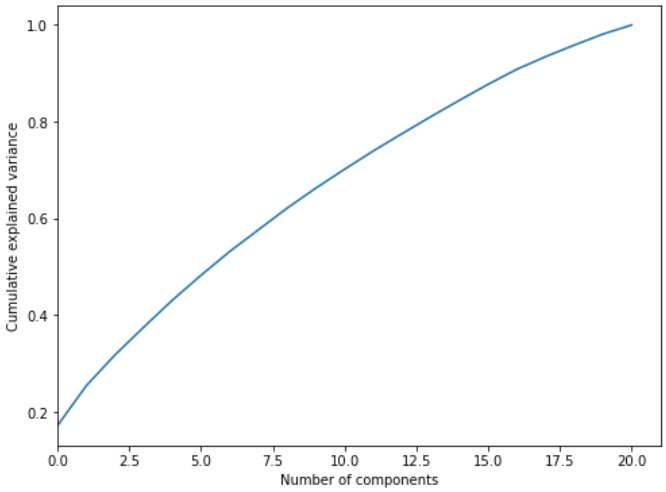


Fig. 4 Cumulative explained variance (Photo credit: Original)

Based on Fig. 4 and Table 3, the dimension is reduced to 17 while reserving more than 90% of the original information.

3.3. Results

After calculation, the result is presented in Table 4. The visualization is shown in Fig. 5.

Table 4. Results

Models	A: train / test (%, 4sf)	P (%, 2sf)	R (%, 2sf)	F1 (%, 2sf)
	Before PCA			
LR	85.24 / 84.23	86	85	85
SVM	86.91/ 86.52	**88**	**87**	**87**
KNN	82.63 / 80.32	83	80	80
DT	86.39 / 84.45	86	84	84
RF	87.18 / 86.13	**88**	86	86
XGB	88.53 / 87.29	**88**	**87**	**87**
	After PCA			
LR	84.97 / 85.30**	86	85	85
SVM	86.91 / 86.51	**88**	**87**	**87**
KNN	82.62 / 80.31	83	80	80
DT	86.33 / 84.88*	87	85	85
RF	87.09 / 86.33*	**88**	86	86
XGB	87.40 / 87.05	**88**	**87**	**87**

*: increase in accuracy after PCA;
**: largest increase in accuracy in PCA;
Bold: highest value in the category

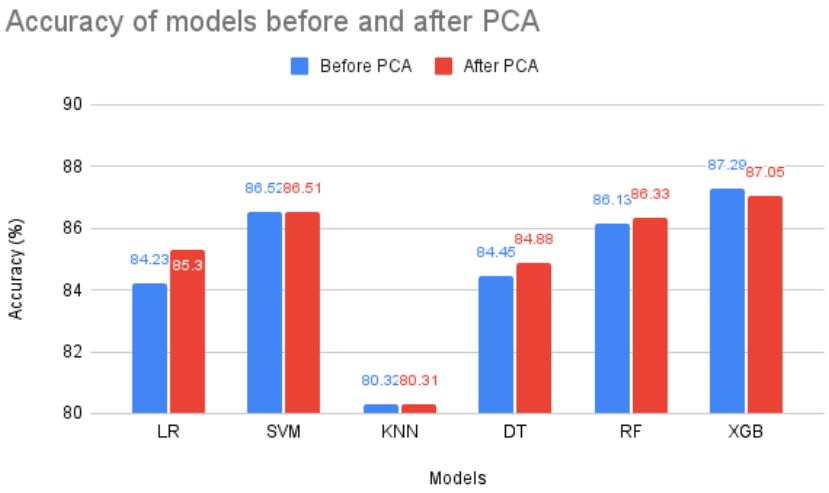


Fig. 5 Comparison of accuracy (Photo credit: Original)

As shown in Table 4 and Figure 5, XGB achieves the greatest accuracy (87.29%), precision (88%), and sensitivity (87%) for both training and test samples, regardless of whether PCA is conducted.

In addition, it is observed that PCA reduces the accuracy of all the models in describing the training data but increases the accuracy of prediction for LR, DT, and RF models. This method has little effect on the accuracy of SVM and KNN models. It has the largest effect on LR model, which increases the prediction accuracy by 1.27%.

4. Conclusion

This study aims to find the best model for diabetes prediction using the common health indicators gathered from health checks and investigate the effectiveness of PCA in improving these models. Based on the results, we can conclude that, with the dataset given, extreme gradient boosting is the best method for diabetes prediction.

It has also been shown that the effectiveness of PCA in enhancing model quality is quite limited. The author suggests that the reason behind this is that the dimension of the dataset used (21) is relatively low compared to the total number of samples (253,680). Thus, the test error produced by non-associated and inter-correlated features is non-significant, limiting the potential of PCA.

4.1. Limitation Analysis

However, this study has several limitations.

Firstly, not all the major causes of diabetes are included. Statistics show that almost half of first-degree relatives of diabetes mellitus type 2 patients may develop the same disease, whereas the incidence rate in the general population is only 6%. Thus, heritable genetic correlation, among others, should be an important reference in diabetes risk assessment [17]. However, no feature variable in the dataset provides information regarding family history.

Secondly, the survey's participants are all American residents, preventing the samples in the dataset from being globally representative. Since no study has proven that race and residential environment have no effect on susceptibility to diabetes, the models built in this study may not reach the same accuracy when applied to different regions and ethnic groups.

4.2. Future Exploration

In the future, the study may consider using other diverse datasets to evaluate the generality and applicability of the models obtained. New models can also be developed and integrated with existing ones to generate collectively strong models that can be extensively used.

Moreover, other feature selection and dimensionality reduction methods may be applied to this dataset to further optimize the model quality. The effectiveness of PCA may be further assessed on high-dimensional datasets instead.

References

[1] Sisodia, D., & Sisodia, D. S. (2018). Prediction of Diabetes using Classification Algorithms. Procedia Computer Science, 132, 1578-1585. https://doi.org/10.1016/j.procs.2018.05.122

[2] Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., & Tang, H. (2018). Predicting Diabetes Mellitus With Machine Learning Techniques. Frontiers in Genetics, 9. https://doi.org/10.3389/fgene.2018.00515

[3] Pradhan, N., Rani, G., Dhaka, V. S., & Poonia, R. C. (2020). Diabetes prediction using artificial neural network. Deep Learning Techniques for Biomedical and Health Informatics, 327-339. https://doi.org/10.1016/B978-0- 12-819061-6.00014-8

[4] El_Jerjawi, N. S., & Abu-Naser, S. S. (2018). Diabetes prediction using artificial neural network. International Journal of Advanced Science and Technology, Vol.121 (2018), pp. 55-64 http://dx.doi.org/10.14257/ijast.2018.121.05

[5] Xie Z, Nikolayeva O, Luo J, Li D. Building Risk Prediction Models for Type 2 Diabetes Using Machine Learning Techniques. Prev Chronic Dis 2019;16:190109. DOI: http://dx.doi.org/10.5888/pcd16.190109

[6] Zhu, C., Idemudia, C. U., & Feng, W. (2019). Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques. Informatics in Medicine Unlocked, 17, 100179. https://doi.org/10.1016/j.imu.2019.100179 \

[7] Choubey Kumar Dilip *, Kumar Manish, Shukla Vaibhav, Tripathi Sudhakar & Dhandhania Kumar Vinay, Comparative Analysis of Classification Methods with PCA and LDA for Diabetes, Current Diabetes Reviews 2020; 16(8). https://dx.doi.org/10.2174/1573399816666200123124008'

[8] Alexteboul. "Diabetes Health Indicators Dataset Notebook." Kaggle, Kaggle, 10 Mar. 2022, https://www.kaggle.com/code/alexteboul/diabetes-health-indicators-dataset-notebook.

[9] Collins, G.S., Mallett, S., Omar, O. et al. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. BMC Med 9, 103 (2011). https://doi.org/10.1186/1741-7015-9-103

[10] Hosmer Jr, D. W., Lemeshow, S., Sturdivant, R. X. (2013). Applied logistic regression (Vol. 398). John Wiley Sons.

[11] Noble, W. What is a support vector machine?. Nat Biotechnol 24, 1565–1567 (2006). https://doi.org/10.1038/nbt1206-1565 [13] ication. In: Meersman, R., Tari, Z., Schmidt, D.C. (eds) On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE. OTM 2003. Lecture Notes in Computer Science, vol 2888. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-39964-3_62 13

[12] Guo, G., Wang, H., Bell, D., Bi, Y., Greer, K. (2003). KNN Model-Based Approach in Classification. In: Meersman, R., Tari, Z., Schmidt, D.C. (eds) On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE. OTM 2003. Lecture Notes in Computer Science, vol 2888. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-39964-3_62

[13] SONG, Y., & LU, Y. (2015). Decision tree methods: applications for classification and prediction. Shanghai Archives of Psychiatry, 27(2), 130-135. https://doi.org/10.11919/j.issn.1002-0829.215044

[14] Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324

[15] Chen, T. & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, 785–794. https://doi.org/10.1145/2939672.2939785

[16] James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). Linear Model Selection and Regularization. In: An Introduction to Statistical Learning. Springer Texts in Statistics, vol 103. Springer, New York, NY. https://doi.org/10.1007/978- 1-4614-7138-7_6

[17] Wu, Y., Ding, Y., Tanaka, Y., & Zhang, W. (2014). Risk factors contributing to type 2 diabetes and recent advances in the treatment and prevention. International journal of medical sciences, 11(11), 1185–1200. https://doi.org/10.7150/ijms.10001