PREDICTION OF DIABETES (SUGAR) USING MACHINE LEARNING TECHNIQUES

Siddamma C.M[1], Sangamitra B.K[2]

[1]Assistant Professor, Department of Internet of Things (IoT)[1], Malla Reddy Engineering College & Management Sciences, Medchal
[2]Assistant Professor, Department of Computer science Engineering, Malla Reddy Engineering College & Management Sciences, Medchal

---***---

Abstract - The aim of data gathering is to derive valuable data from broad data set We need to consider the role before learning what Diabetes(‘sugar’) is Insulin is used as a "gateway" to unlock. The corporeal structures allowing our bodies to use glucose for energy Insulin regulates our body 's intake of glucose Diabetes(‘sugar’) is a disorder in which blood glucose levels climb Predominantly physical and comical examination has identified Diabetes(‘sugar’) but it does not provide conclusive outcomes To overcome this constraint we allow disease prediction using numerous Data gathering algorithms for Diabetes(‘sugar’) mellitus prediction and diagnosis.

Information mining likewise entitled information revelation in databanks in software engineering the improvement of finding invigorating and significant examples and relationship in immense volumes of information He field consolidates apparatuses from insights and computerized reasoning for example neural organizations and AI with information base administration to examine enormous advanced assortments known as informational collections Information mining is Heavily used in business protection Study in asset management research cosmology medication and government security recognition of lawbreakers and psychological oppressors.

Random Forest XGBoost and Logistic Regression are the key knowledge mining calculations discussed in this article the knowledge index chosen for exploratory replication focuses on the Pima Indian Diabetic Collection from the University of “California “Irvine UCI Machine Learning Data Sets Repository Prelude.

***Key Words*: K-NN, SRS, SVM, PNN, BLR, PLS-DA, Positive precision, XGBOOST.**

1.INTRODUCTION

Data gathering additionally entitled information disclosure in databanks in software engineering the improvement of finding animating and important examples and relationship in colossal volumes of information He field joins devices from measurements and manmade reasoning for example neural organizations and AI with information base administration to investigate enormous computerized assortments known as informational indexes Information mining is Heavily used in business protection Study in fund management cosmology medication and government security discovery of hoodlums and fear based oppressors. [2] The disease's estimation assumes a significant function in information mining There are various kinds of ailments anticipated in information mining to be specific Hepatitis 'Lung' ' Cancer' 'Liver' issue Breast malignant growth Thyroid malady Diabetes(‘sugar’) and so forth… This paper dissects the Diabetes(‘sugar’) expectations There are essentially four kinds of Diabetes(‘sugar’) Mellitus They are group1 group2 Gestational Diabetes(‘sugar’) intrinsic Diabetes(‘sugar’).

In group 1 Diabetes(‘sugar’) the human doesn't create insulin It is typically analysed in kids and youthful grownups and was recently known as adolescent Diabetes(‘sugar’). Just 5 per cent of people with Diabetes (‘sugar’) have the condition in this manner.

In group 2 the most prevalent form of Diabetes(‘sugar’) is type 2 Diabetes(‘sugar’). The body should not adequately use insulin in this phase. This is known as “insulin resistance”.

The Congenital Diabetes(‘sugar’)(‘sugar’) is caused because of hereditary deformities of insulin emission cystic fibrosis related Diabetes(‘sugar’)(‘sugar’) steroid Diabetes(‘sugar’)(‘sugar’) instigated by high dosages of glucocorticoids If it is untreated or inappropriately oversaw Diabetes(‘sugar’)(‘sugar’) can bring about an assortment of inconveniences including coronary episode stroke kidney disappointment visual impairment issues with erection feebleness and removal Keeping your circulatory strain and blood glucose sugar at target will assist with dodging Diabetes(‘sugar’)(‘sugar’) confusions For this it ought to be analyzed as right on time as conceivable to give appropriate treatment.

The greatest value of digital management is that hospitals are constantly storing and tracking a large data storage of previous patient history with multiple comparisons. Such patient data enables physicians to discover multiple variations in the data collection. Diseases may be collected, forecast and diagnosed using the designs found in data sets.

2. PROPOSED SYSTEM

Different numbers of methods of data collection exist. One technique for categorizing multiple methods of data collection is based on their ability to function. Regression is a technique for mathematics that is mostly used for numerical estimation. Gathering returns a collection of documents to their propensities. In a sequence set of data / information, the sequential pattern mechanism searches for repeated sub sequences where a sequence logs an order of events. To conclude, a compact definition is to be made for a subset of data.

Stage 1: A classifier representing a pre-determined set of data classes or definitions is constructed. This is the lifelong journey (or training phase), where a classification algorithm constructs the algorithm by evaluating or "learning from" its training set consisting of database tuples and their corresponding class labels. It is assumed that each tuple belongs to a predefined class called the attribute mark class.Since each coaching tuple 's class mark is given, the phase is also known as directed classification. In comparison, the first step can be seen as studying a mapping or function, y = f (X), which can predict the corresponding class mark y of a given tuple X. This mapping is usually expressed in the form of rules of grouping, decision trees, or theoretical formulae.

Step 2: For grouping, the model is used. Firstly, the classifier 's predictive precision is calculated. If we had to use the training plan to gauge the trained model 's accuracy.

Benefits:

A bonus is that such training approaches, such as what you see in arbitrary forests, offer a form of global feature selection, as any flaws of any specific case are also compensated by selecting from a random subset of features from version to model through selfish, local transfer learning (e.g. knowledge gain).

Providing detailed outcomes for Diabetes(‘sugar’) prediction in the proposed method.

3. MODULES

1. Admin

2. Patient Collection of data/information

3. Data gathering Algorithm

4. Prediction

Modules Description:

1. Admin: Admin gathers information about the customer. And submit classified intelligence gathering on data / information compilation for patients.

2. Patient Collection of data/information:It is used to load data / information obtained by the patient consisting of:

The checklist for data collection consists of 17 sub-questions such as surname, age, weight, physical exercise, urination, water intake, diet, systolic blood pressure, hypertension, exhaustion, blurred vision, wound recovery, sleepy / nauseous, abrupt weight loss, genetics, glucose level and diabetic mellitude.

Max data set weight is 255, with 17 attributes. All information gathered is stored in the Excel spreadsheet file (xls) format. Is being used in the research data collection using different classifiers to forecast Diabetes(‘sugar’) mellitus.

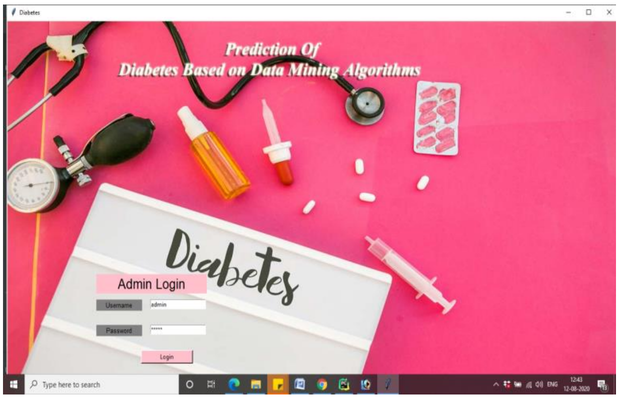
Data gatheringAlgorithm:

We monitor the effectiveness of C4.5, SVM, k-NN, PNN, BLR, MLR, PLS-DA, PLS-LDA, k-means and Apriori and then compare the success of data collection algorithms based on computation time, accuracy score, data evaluated using 10-fold Cross Validation Failure Ranking, True Positive, True Negative, False “+” and False “-”, validation and accuracy of bootstrap. A typical confusion matrix is furthermore displayed for quick check.

4. Prediction

The algorithm's output is measured using the Overall Consistency and Spontaneous Accuracy equations. The True “+” and True “-”, False “+” and False “-” conditions for evaluating the calculation are taken here.

4. INTERPRETATION OF RESULTS



Prediction Of
Diabetes Based on Data Mining Algorithms

Admin Login

Username: admin

Password: *****

Login

Fig-1. Admin Login

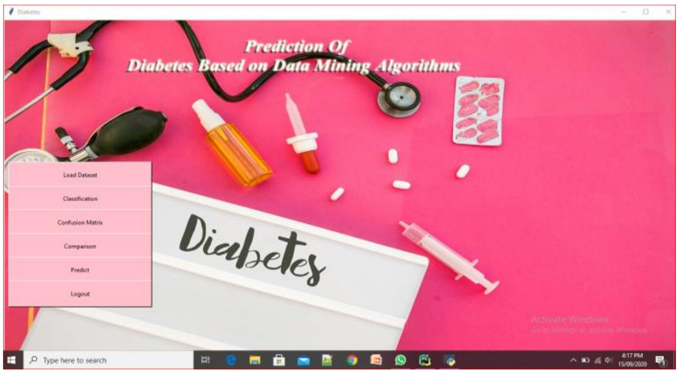
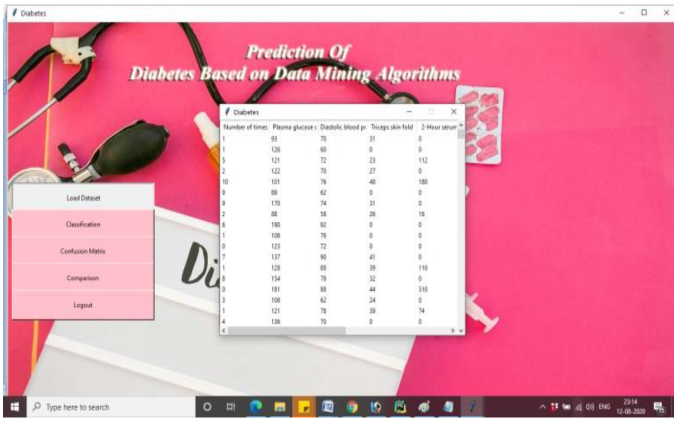
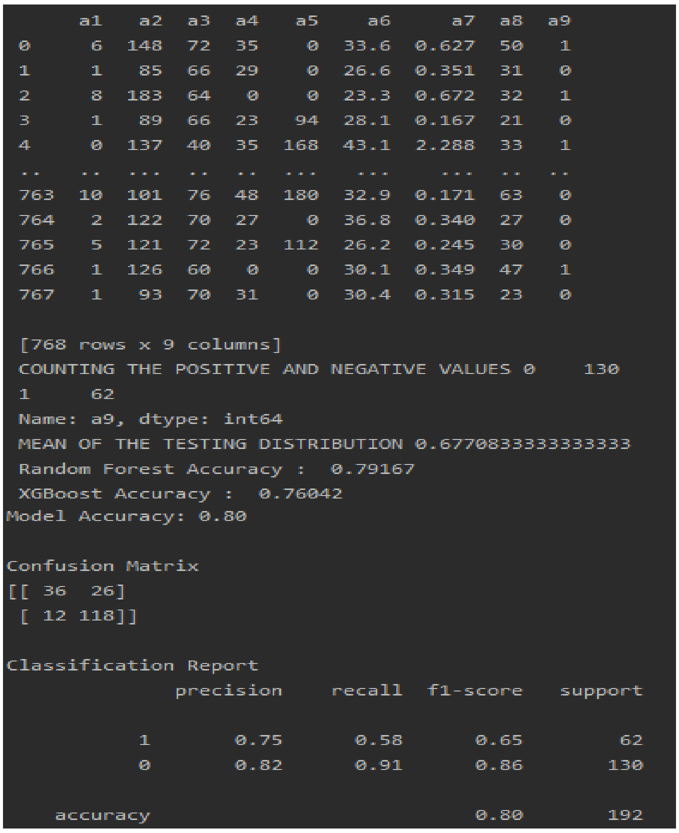


Fig-2.Menu



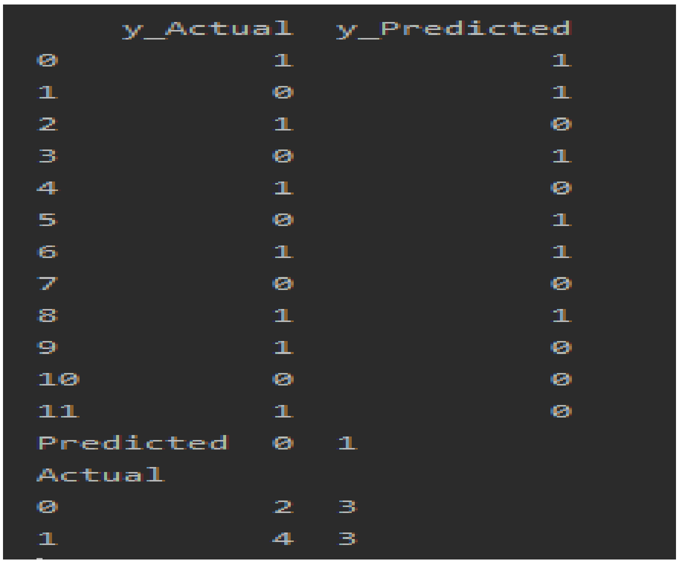
Number of times	Plasma glucose c	Diastolic blood pr	Triceps skin fold	2-Hour serum
1	93	70	31	0
1	126	60	0	0
5	121	72	23	112
2	122	70	27	0
10	101	76	48	180
9	89	62	0	0
9	170	74	31	0
2	88	58	26	16
6	190	92	0	0
1	106	76	0	0
0	123	72	0	0
7	137	90	41	0
1	128	88	39	110
8	154	78	32	0
0	181	88	44	510
3	108	62	24	0
1	121	78	39	74
4	136	70	0	0

Fig-3. Load Collection of data/information



```
      a1   a2  a3  a4   a5    a6     a7  a8  a9
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
..    ..  ...  ..  ..  ...   ...    ...  ..  ..
763   10  101  76  48  180  32.9  0.171  63   0
764    2  122  70  27    0  36.8  0.340  27   0
765    5  121  72  23  112  26.2  0.245  30   0
766    1  126  60   0    0  30.1  0.349  47   1
767    1   93  70  31    0  30.4  0.315  23   0

 [768 rows x 9 columns]
 COUNTING THE POSITIVE AND NEGATIVE VALUES 0    130
 1     62
 Name: a9, dtype: int64
 MEAN OF THE TESTING DISTRIBUTION 0.6770833333333333
 Random Forest Accuracy :  0.79167
 XGBoost Accuracy :  0.76042
Model Accuracy: 0.80

Confusion Matrix
[[ 36  26]
 [ 12 118]]

Classification Report
              precision    recall  f1-score   support

           1       0.75      0.58      0.65        62
           0       0.82      0.91      0.86       130

    accuracy                           0.80       192
```

Fig-4.Classification



```
        y_Actual  y_Predicted
0              1            1
1              0            1
2              1            0
3              0            1
4              1            0
5              0            1
6              1            1
7              0            0
8              1            1
9              1            0
10             0            0
11             1            0
Predicted      0    1
Actual
0              2    3
1              4    3
```

Fig-5.Confusion Matrix

This interface is used to show the Matrix of Uncertainty. An uncertainty matrix is a table frequently used to define the results of a classification model (or "classifier") on a collection of test data that are considered to be true values. It enables the visualization of an algorithm's output.

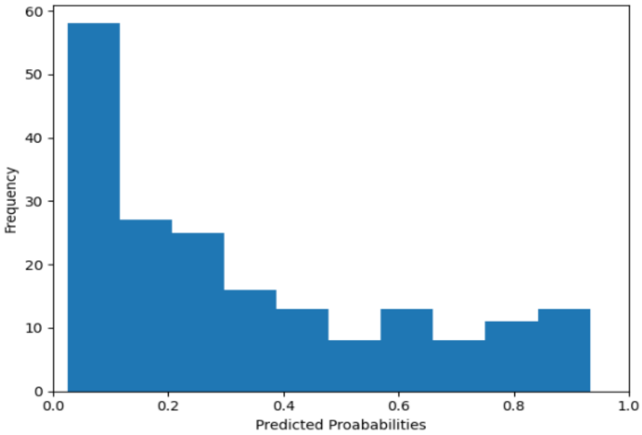


Fig-6.Prediction

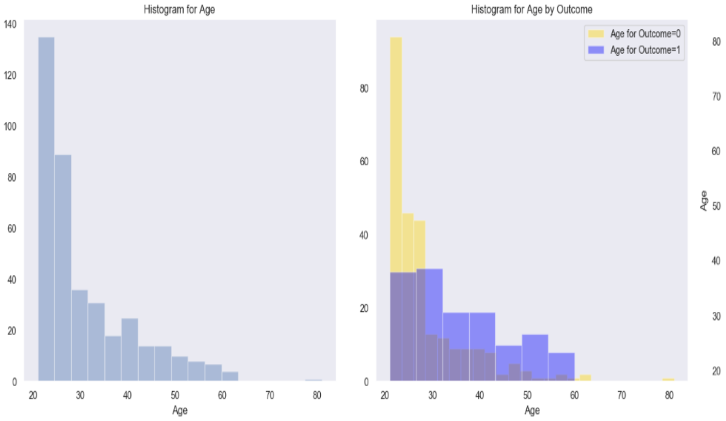


Fig-7. Histogram for age attribute

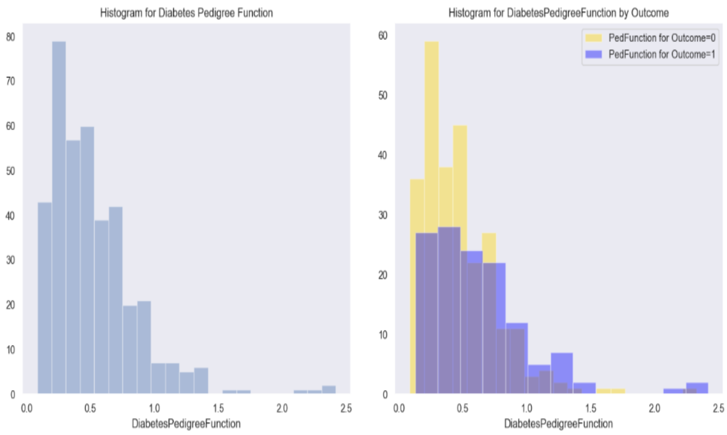


Fig-8.Histograms for Diabetes ('sugar') pedigree function attribute

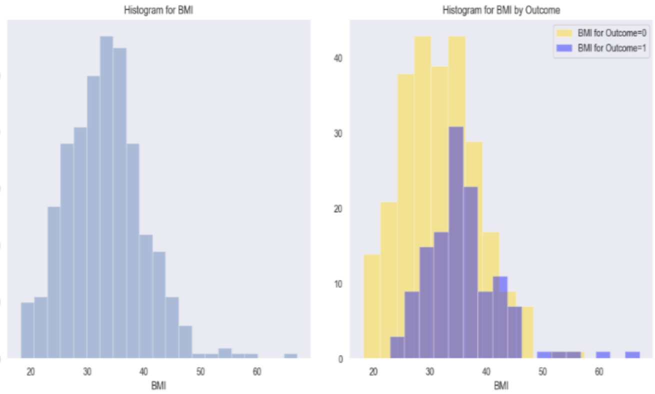


Fig-9. Histogram for BMI Attribute

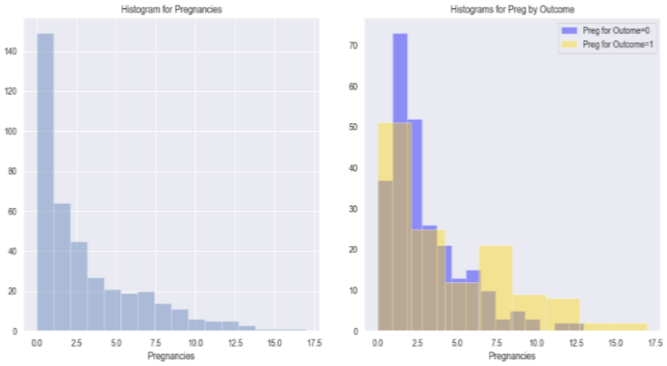


Fig-10.Histogram for Pregnancy attribute

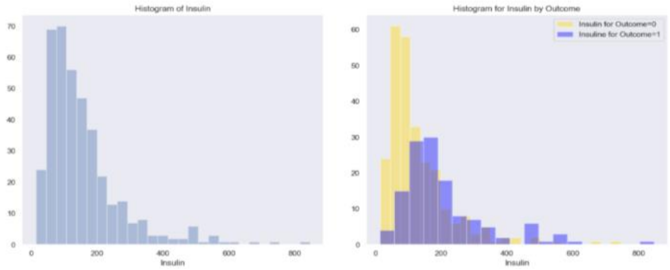


Fig-11.Histogram for Insulin

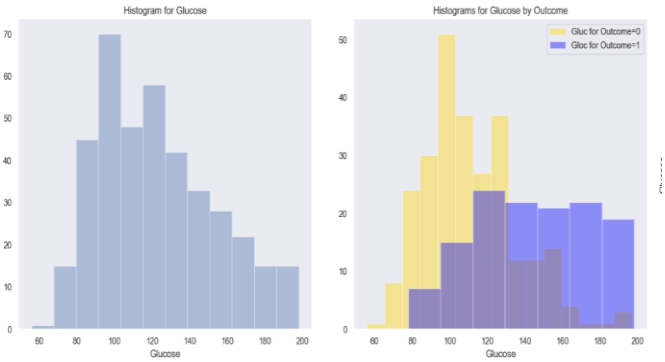


Fig-12.Histogram for glucose Attribute

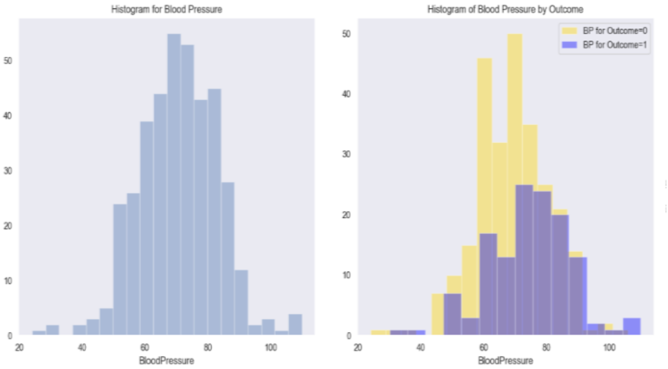


Fig-13.Histogram for Blood pressure Attribute

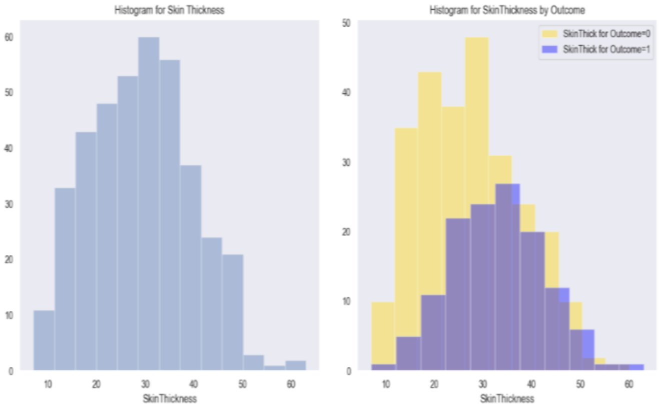


Fig-14.Histogram for Skin thickness

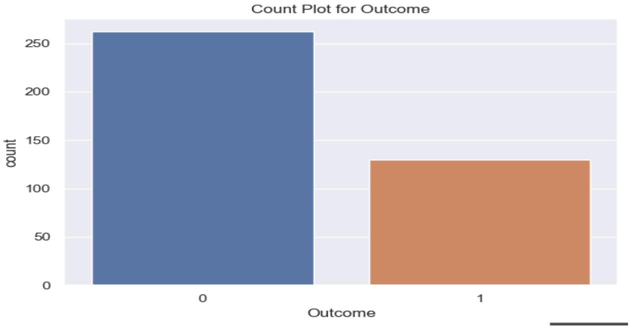


Fig-15.Count plot for outcome of Diabetes('sugar') patients

5. CONCLUSIONS

Machine learning really does have the great potential to revolutionize the estimation of Diabetes('sugar') risk by sophisticated statistical techniques and the provision of vast volumes of data / information gathering threats for epidemiological and genetic Diabetes('sugar'). The secret to recovery is the diagnosis of Diabetes('sugar') in its early phases.

This thesis identified an interface to machine learning to anticipate the Diabetes('sugar') levels. The approach can also help clinicians develop a reliable and efficient instrument to help consumers make informed choices about the state of the disease at the clinician 's table.

REFERENCES

1. Alghamdi M., Al-Mallah M., Keteyian S., Brawner C., Ehrman J., Sakr S. (2017). Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: the henry ford exercise testing (FIT) project. PLoS One 12:e0179805. 10.1371/journal.pone.0179805
2. American Diabetes Association (2012). Diagnosis and classification of diabetes mellitus. Diabetes Care 35(Suppl. 1), S64–S71. 10.2337/dc12-s064
3. Bengio Y., Grandvalet Y. (2005). Bias in Estimating the Variance of K -Fold Cross-Validation. New York, NY: Springer, 75–95. 10.1007/0-387-24555-3_5
4. Breiman L. (2001). Random forest. Mach. Learn. 45 5–32. 10.1023/A:1010933404324
5. Chen X. X., Tang H., Li W. C., Wu H., Chen W., Ding H., et al. (2016). Identification of bacterial cell wall lyases via pseudo amino acid composition. Biomed. Res. Int. 2016:1654623. 10.1155/2016/1654623
6. Cox M. E., Edelman D. (2009). Tests for screening and diagnosis of type 2 diabetes. Clin. Diabetes 27 132–138. 10.2337/diaclin.27.4.132
7. Duygu ç., Esin D. (2011). An automatic diabetes diagnosis system based on LDA-wavelet support vector machine classifier. Expert Syst. Appl. 38 8311–8315.
8. Friedl M. A., Brodley C. E. (1997). Decision tree classification of land cover from remotely sensed data. Remote Sens. Environ. 61 399–409
9. Georga E. I., Protopappas V. C., Ardigo D., Marina M., Zavaroni I., Polyzos D., et al. (2013). Multivariate prediction of subcutaneous glucose concentration in type 1 diabetes patients based on support vector regression. IEEE J. Biomed. Health Inform. 17 71–81. 10.1109/TITB.2012.2219876
10. Habibi S., Ahmadi M., Alizadeh S. (2015). Type 2 diabetes mellitus screening and risk factors using decision tree: results of data mining. Glob. J. Health Sci. 7 304–310. 10.5539/gjhs.v7n5p304
11. Han L., Luo S., Yu J., Pan L., Chen S. (2015). Rule extraction from support vector machines using ensemble learning approach: an application for diagnosis of diabetes. IEEE J. Biomed. Health Inform. 19 728–734. 10.1109/JBHI.2014.2325615.
12. Iancu I., Mota M., Iancu E. (2008). "Method for the analysing of blood glucose dynamics in diabetes mellitus patients," in Proceedings of the 2008 IEEE International Conference on Automation, Quality and Testing, Robotics, Cluj-Napoca: 10.1109/AQTR.2008.4588883
13. Jackson D. A. (1993). Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches. Ecology 74 2204–2214. 10.2307/1939574
14. Jegan C. (2014). Classification of diabetes disease using support vector machine. Microcomput. Dev. 3 1797–1801.
15. Jia C., Zuo Y., Zou Q. (2018). O-GlcNAcPRED-II: an integrated classification algorithm for identifying O-GlcNAcylation sites based on fuzzy undersampling and a K-means PCA oversampling technique. Bioinformatics 34 2029–2036. 10.1093/bioinformatics/bty039
16. Jiang Y., Zhou Z. H. (2004). Editing training data for kNN classifiers with neural network ensemble. Lect. Notes Comput. Sci. 3173 356–361. 10.1007/978-3-540-28647-9_60
17. Jolliffe I. T. (1998). "Principal components analysis," in Proceedings of the International Conference on Document Analysis and Recognition (Heidelberg: Springer;).
18. Kavakiotis I., Tsave O., Salifoglou A., Maglaveras N., Vlahavas I., Chouvarda I. (2017). Machine learning and data mining methods in diabetes research. Comput. Struct. Biotechnol. J. 15 104–116. 10.1016/j.csbj.2016.12.005
19. Kim J. H. (2009). Estimating classification error rate: repeated cross-validation, repeated hold-out and bootstrap. Comput. Stat. Data Anal. 53 3735–3745. 10.1016/j.csda.2009.04.009
20. Kohabi R. (1996). "Scaling up the accuracy of naive-bayes classifiers : a decision-tree hybrid," in Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, Portland, OR
21. Kohavi R. (1995). "A study of cross-validation and bootstrap for accuracy estimation and model selection," in Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montreal

22. Krasteva A., Panov V., Krasteva A., Kisselova A., Krastev Z. (2011). Oral cavity and systemic diseases—Diabetes Mellitus. Biotechnol. Biotechnol. Equip. 25 2183–2186. 10.5504/BBEQ.2011.0022

23. Lee B. J., Kim J. Y. (2016). Identification of type 2 diabetes risk factors using phenotypes consisting of anthropometry and triglycerides based on machine learning. IEEE J. Biomed. Health Inform. 20 39–46. 10.1109/JBHI.2015.2396520 .

24. Li B. Q., Zheng L. L., Feng K. Y., Hu L. L., Huang G. H., Chen L. (2016). Prediction of linear B-cell epitopes with mRMR feature selection and analysis. Curr. Bioinform. 11 22–31. 0.2174/1574893611666151119215131

25. Liao Z., Ju Y., Zou Q. (2016). Prediction of G protein-coupled receptors with SVM-Prot features and random forest. Scientifica 2016:8309253. 10.1155/2016/8309253

26. Liao Z. J., Wan S., He Y., Zou Q. (2018). Classification of small GTPases with hybrid protein features and advanced machine learning techniques. Curr. Bioinform. 13 492–500. 10.2174/1574893612666171121162552

27. Liaw A., Wiener M. (2002). Classification and regression by randomforest. R. News 2 18–22.

28. Lin C., Chen W., Qiu C., Wu Y., Krishnan S., Zou Q. (2014). LibD3C: ensemble classifiers with a clustering and dynamic selection strategy. Neurocomputing 123 424–435. 10.1016/j.neucom.2013.08.004

29. Lonappan A., Bindu G., Thomas V., Jacob J., Rajasekaran C., Mathew K. T. (2007). Diagnosis of diabetes mellitus using microwaves. J. Electromagnet. Wave. 21 1393–1401. 10.1163/156939307783239429 .

30. Mukai Y., Tanaka H., Yoshizawa M., Oura O., Sasaki T., Ikeda M. (2012). A computational identification method for GPI-anchored proteins by artificial neural network. Curr. Bioinform. 7 125–131. 10.2174/157489312800604390

31. Ozcift A., Gulten A. (2011). Classifier ensemble construction with rotation forest to improve medical diagnosis performance of machine learning algorithms. Comput. Methods Programs Biomed. 104