Machine Learning based approach for Diabetes Prediction

Juganta Dutta, *Sonali Mondal, Biswajit Das

Department of Computer Science, Arunachal University of Studies, Namsai, Arunachal Pradesh, India

*Corresponding author E-mail: sonalimondal20387@gmail.com , biswajit195313@gmail.com

---***---

Abstract - Diabetes is an illness brought on by an excessive amount of glucose within the body. Ignorance of diabetes is no longer acceptable. If neglected, it may also result in more severe health concerns for a person, such as heart-related problems, renal problems, blood pressure, eye damage, and effects on other body organs. Insulin hormone is affected, which leads to abnormal crab metabolism and elevates blood sugar levels. According to the World Health Organization, 422 million people worldwide suffer with diabetes. low- and middle-class people being disproportionately affected. The condition is caused by the body producing insufficient amounts of insulin. Additionally, this might reach 490 billion by 2030. To benefit from this challenging job, we may apply ensemble techniques and system learning for classification on this image to forecast whether diabetes will be present in a dataset. When comparing one version to another, the accuracy varies depending on the model. The assignment provides the accurate or improved accuracy version, indicating that the model can effectively predict diabetes. Our findings demonstrate that random forested areas outperformed other system mastery strategies in terms of accuracy.

Keywords: Classification Algorithms, Supervised Learning, Unsupervised Learning, Random forest

1. INTRODUCTION

Diabetes mellitus, normally called diabetes, is a chronic metabolic sickness characterized through improved blood sugar ties over a prolonged period. It affects tens of million cutting-edge humans international and a prolonged period. It affects tens of millions cutting-edge humans international and poses enormous demanding situations to public fitness systems globally in keeping with the global according to the Diabetes Federation, 463 million persons (aged 20 to 79) have diabetes in 2019, with projections estimating this number to upward thrust to seven hundred million via 2045. With such alarming facts, early detection and prevention techniques are paramount to mitigation the burden cutting-edge this ailment on people and society[1].

In latest years, the sphere present day gadget modern day has emerged as a effective tool for healthcare specialists to expect, diagnose, and manage numerous clinical conditions, along with diabetes. By way of leveraging superior algorithms and computational techniques, device state of the art models can analyze giant quantities ultra-modern affected person facts to identify patterns, hazard factors and predictive markers related to diabetes onset, those models provide the potential to revolutionize traditional healthcare practices by means of enabling personalized, records-driven interventions tailored to individual patient needs[3].

The goal cutting edge this paper is to provide a comprehensive evaluation cutting edge the modern day techniques in predicting diabetes risk the usage of gadget present day algorithms. We are able to discover the numerous information assets, function choice strategies, model architectures, and assessment metrics hired in present research. Moreover, we are able to discuss the challenges, obstacles, and destiny instructions modern day making use of system contemporary for diabetes prediction.

2. LITERATURE REVIEW

The provided work's analysis has an impact on a variety of healthcare datasets, where evaluation and prediction have been carried out utilizing a range of methodologies. Many academics have created and tested a wide range of prediction models using various iterations of information algorithms for machine learning, mining techniques, or mixes of those methods. This gadget predicts the kind of diabetes and its associated risks. The device is highly cost-effective for any healthcare institution and is entirely Hadoop based. [12]Arora(2015) examined hidden patterns in the diabetes dataset using the category approach. This model makes use of selection timber and naïve Bayes.

Finding and computing the percentages of accuracy, sensitivity, and specificity of various classification techniques is the aim of Choubey et al. [13].. They also attempted to evaluate and compare the outcomes of several classification techniques using WEKA. Additionally, the study evaluates how well the same classifiers work when used with other programs, while keeping the same characteristics (i.e., accuracy, sensitivity, and specificity), like Matlab and Rapidminer.

The JRIP, Jgraft, and BayesNet algorithms were employed. Jgraft has the highest accuracy (81.3%), sensitivity (59.7%), and specificity (81.4%), according the results. Furthermore, it was found that WEKA outperforms Matlab and Rapidminner.

After applying the resample filter on the diabetic dataset, Choubey et al. [14] concentrate on using the CART decision tree technique. In order to obtain higher accuracy rates, the author places focus on the class imbalance issue and the necessity of addressing it before using any algorithms. Datasets having dichotomous values—that is, when the class variable has two possible outcomes—often exhibit class imbalances. The accuracy of the prediction model will increase if these imbalances are identified early in the data preparation step and are easily addressed.

The taxonomy displayed in the image below aids in the system's learning of possible diabetes prediction algorithms. The challenge in selecting a device learning algorithm is matching aspects of the data to be found only based on techniques that are currently in use. This article discusses the taxonomy of device learning algorithms. System learning can be classified into three categories: semi-supervised learning, unsupervised learning, and supervised learning.

In Procedia Computer Science 165 (2019), 292–299, Aishwarya Mujumdar et al.

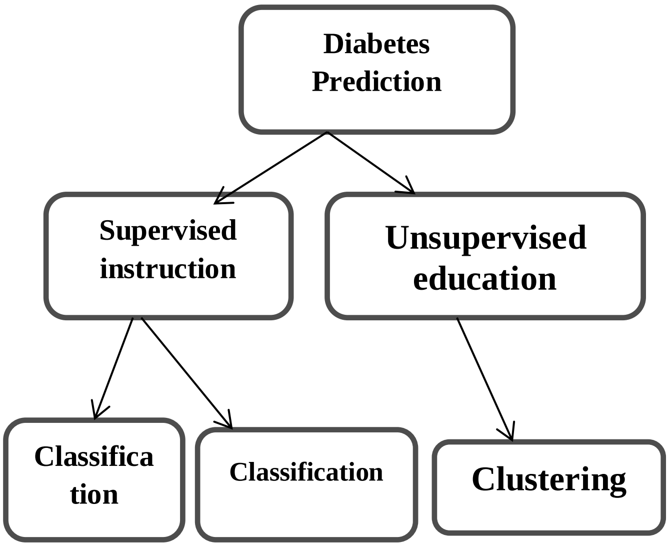


Figure 1

A. The Supervised Acquaintance/Predictive Styles Algorithms for supervised learning are used to create predictive styles. Using various values found in the dataset, a predictive algorithm forecasts missing values. With a given set of input records and an extra fixed set of output records, the supervised learning method creates a model that enable useful predictions about the reaction to fresh datasets. Choice Trees, Bayesian techniques, Supervised learning techniques include artificial neural networks, instance-based mastering, and ensemble methods. These are really well-liked techniques for learning devices.

B. Using unsupervised learning and descriptive models learning unsupervised study methods have led to the evolution of descriptive models. Although the output is unclear, we are aware of a set of inputs on this version. Transactional data is primarily utilized for learning without supervision. This technique includes clustering techniques like ok-way clustering and ok-Medians clustering.

C. Studying under a semi-supervised environment every classified and unlabeled statistic in the education dataset is used in the semi-supervised learning method. Regression approaches are a subset of semi-supervised learning in the classroom.

3. PROPOSED METHODOLOGY

This section will cover the various classifiers that are employed in the system learning to forecast diabetes. We will also go over our suggested approach to improve the accuracy. This paper has employed five unique strategies. The several methods are described below. The output is the gadget that measures the accuracy metrics of fashions. After that, the version can be used for forecasting.

An explanation of the dataset

The dataset was obtained from diabetes.csv on https://github.com/sidhardhan23/main. 769 instances are in this dataset. Determining if the patient has diabetes or not is the primary goal.

The product is reflected in the result. 1 indicates that the individual has diabetes 0 indicates that the individual does not have diabetes.

- 769 items in the range index, 0 to 768

Table 1

1	Pregnancies	Glucose	Blood pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree function	Age	Outcome
2	6	148	72	35	0	33.6	0.627	50	1
3	1	85	66	29	0	26.6	0.351	31	0
4	8	183	64	0	0	23.3	0.672	32	1
5	1	89	66	23	94	28.1	0.167	21	0
6	0	137	40	35	168	43.1	2.288	33	1
7	5	116	74	0	0	25.6	0.201	30	0
8	3	78	50	32	88	31	0.248	26	1

- This dataset contains 769 cases with 9 features each.

<class ‘pandas.core.frame.dataframe’>

Range index: 769 entries, 0 to 768

Datatypes: float64(2) , int64(7)

Memory usage: 23.3 kb

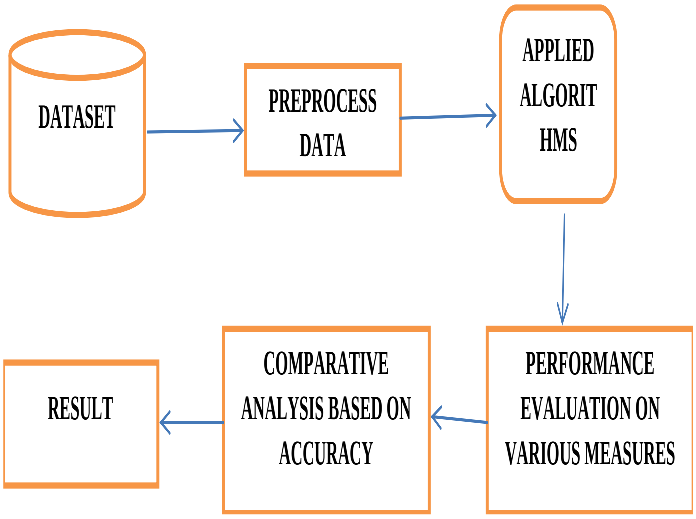

DATASET
PREPROCESS DATA
APPLIED ALGORITHMS
PERFORMANCE EVALUATION ON VARIOUS MEASURES
COMPARATIVE ANALYSIS BASED ON ACCURACY
RESULT

Figure 2-suggested model illustration

RESULT AND DISCUSSION

It is easy to observe that our outcome price is not entirely correlated with any one function. A small number of attributes have poor correlations with the result value, while a small number have excellent permits. Analyze the plots. Furthermore, it verifies the necessity of scalability by demonstrating how every feature and label is distributed among extraordinary levels. Next, it basically approaches that each discrete bar you see is a true independent variable. We would want to deal with these explicit factors before using device learning. We have two training labels for our outcome: 0 for no illness and 1 for a disease.

In the plot above, the y-axis represents education and check set precision, and the location of n neighbors is represented by the x-axis. Considering that the prediction at the schooling set is flawless if we select the sole nearest neighbor. However, the accuracy of education decreases when larger friends are included, suggesting that using the nearest unmarried neighbor leads to an overly complex version. Somewhere around nine friends is the fantastic performance.

The aforementioned graph indicates that the data is skewed in favor of data points with zero outcome cost, indicating that diabetes was no longer a free gift. The range of those without diabetes is approximately double that of people with the disease.

Both the training and testing processes have excellent accuracy rates, with training processes being 100% accurate. Relevance of Features in Decision Trees The significance of every decision tree is gauged by its feature importance. A value between 0 and 1, which denotes "never used" and "perfect target," corresponds to each attribute

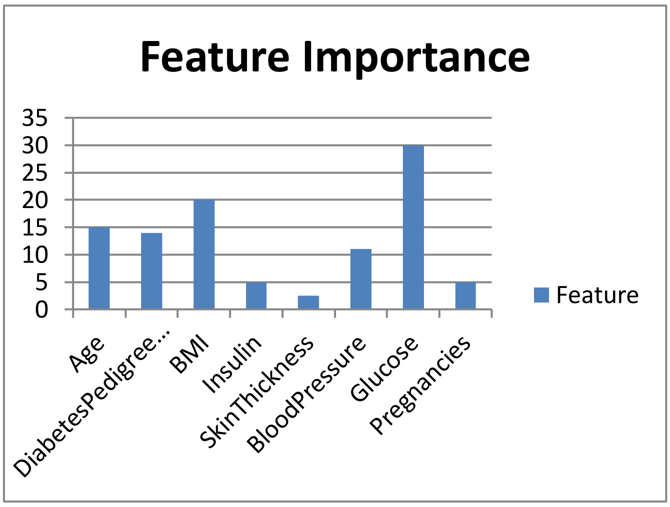
The "glucose" characteristic is the most significant by far. It produces a forest of trees, with each tree having randomly chosen characteristics drawn from the entire set of features.

Table 2-Training accuracy table

Accuracy of Training	1.00
Verifying Correctness	0.769

Figure 3-A feature's significance in Random Forest:

Random forests choose "BMI" as the feature with the second largest feature out of all the data, much like single decision trees do. By altering the distance between the data points and the hyperplane; groups can be successfully divided by the hyperplane. It is determined by a number of kernels. Four cores were tested by me: sigmoid, poly, rbf, and linear.



CONCLUSION

Diagnosing diabetes early is one of the world's most significant medical issues. In this work, we attempted to develop a blood sugar estimation system. This study looked at and assessed five machine learning classification methods using different measures. The John Diabetes database was tested. Results from the experiment verified that the suggested strategy was appropriate in other circumstances and that the decision tree algorithm produced 99% accuracy. By utilizing some machine learning techniques, this study can be enhanced and expanded to provide a more accurate diabetes diagnosis.

REFERENCES

[1] International Journal of Engineering Research & Technology (IJERTISSN: 2278-0181Vol. 9 Issue 09, September-2020

[2]. Arora, R., Suman, 2012. International Journal of Computer Applications 54, 21–25. doi: 10.5120/8626-2492. Bamnote, M.P., G.R., 2014. Create a classification system for diagnosing diabetes using genetic programming.

[3]. Choubey, D.K., Paul, S., Kumar, S., Kumar, S., 2017. Proceedings of the International Conference on Communication and Communication Technology (ICCCS 2016), page 123. 451 → 455.DhomseKanchan B., M.K.M., 2016. Sayfa 5-10. Sharief, A.A., Sheta, A., 2014

[4]. Sisodia, D., Shrivastava, S.K., Jain, R.C., 2010. ISVM is used for facial recognition. Proceedings - 2010 International Conference on Computational Intelligence and Communication Networks, CICN 2010, 554-559doi:10.1109/CICN.2010.109.

[5]. Sisodia, D., Singh, L., Sisodia, S., 2014. 28-30 March 2012, Springer.Pages 1027-1038 Ib.

[6]AishwaryaMujumdar et al. / Procedia Computer Science 165 (2019) 292–299

[7] Gauri D. Kalyankar, Shivananda R. Poojara and Nagaraj V

[8] AyushAnandthiabDivya Shakti, "Diabetes Prediction Based on Personal Lifestyle Indicators," 1st International Conference on Next Generation Computer Technologies, 978-1-4673-6809-4, September 2015.

[9]Dr. B. Nithyathiab V. Ilango, Predictive analytics in healthcare using machine learning tools and techniques, International Conference on Intelligent Computing and Control Systems, 978-1-5386-2745-7, 2017.

[10]https://www.researchgate.net/publication/347091823_Diabetes_Prediction_Using_Machine_Learning

[11] Dr.Saravanakumar NM, Eswari T, Sampath P thiabLavanya S, "Prediction methods for diabetes data

analysis in big data", 2nd International Symposium on Big Data and Cloud Computing, 2015.

[12]. Arora, R., Suman, 2012. Comparative Analysis of Classification Algorithms on Different Datasets using WEKA. International Journal of Computer Applications 54, 21–25. doi:10.5120/8626-2492.

[13] Bamnote, M.P., G.R., 2014. Developing a Classifier via Genetic Programming to Identify Diabetes Mellitus.doi:10.1007/978-3-319-11933-5. Advances in Intelligent Systems and Computing 1, 763–770.

[14] Choubey, D.K., Kumar, S., Paul, S., Kumar, S., 2017 . Using naive bayes and a genetic algorithm for attribute selection, the Pima Indian Diabetes Dataset was classified. This information was published in Communication and Computing Systems: Proceedings of the International Conference on Communication and Computing System (ICCCS 2016), pages 451–455.