Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 8, August : 2024

PREDICTING EARLY-STAGE DIABETES USING MACHINE LEARNING TECHNIQUES

[1] VUTUKURI SUSMITHA, [2] MRS. CH. DEEPTI

[1] PG Scholar in the Department of MCA at QIS College of Engineering & Technology (AUTONOMOUS), Vengamukkapalem, Ongole- 523272, Prakasam Dt., AP., India.

[2] Assistant Professor in the Department of CSE/MCA at QIS College of Engineering & Technology (AUTONOMOUS), Vengamukkapalem, Ongole- 523272, Prakasam Dt., AP., India.

ABSTRACT

As per the report of the World Health Organization (WHO), diabetes has become one of the rapidly expanding chronic diseases that has affected the life of 422 million people all over the world. The number of deaths in Bangladesh due to diabetes has reached 28,065, which is 3.61% of the total deaths of Bangladesh, according to the latest data published by the WHO in 2018. So, we need to be concerned about the risks of diabetes disease. If we cannot take proper steps to diagnose diabetes at an early stage, eventually we have to face serious health issues. In this paper, we have shown the relation of different symptoms and diseases that cause diabetes so that we can help a person to diagnose diabetes at an early stage. Nowadays, machine learning classification approaches are well accepted by researchers for developing disease risk prediction models. Therefore, eleven machine learning classification algorithms such as Logistic Regression (LR), Gaussian Process (GP), Adaptive Boosting (AdaBoost), Decision Tree (DT), K-Nearest Neighbors (KNN), Multilayer Perceptron (MLP), Support Vector Machine (SVM), Bernoulli Naive Bayes (BNB), Bagging Classifier (BC), Random Forest (RF), and Quadratic Discriminant Analysis (QDA) have been used in this study. Among all these machine learning classifiers, Random Forest (RF) classifier has showed the best accuracy of 98%. And its Area Under Curve (AUC) is also, the highest

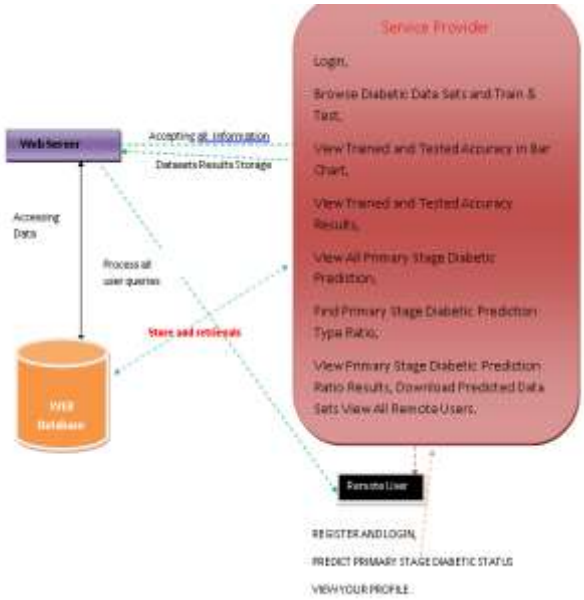
INDEX: world health organization, machine learning classification, early diabetic, prediction

INTRODUCTION:

Diabetes is a chronic condition that develops when the pancreas does not contain enough insulin or when the body does not use the insulin it does produce sufficiently. Insulin is a hormone that regulates blood sugar levels. Type 1 diabetes (also known as insulin-dependent diabetes, juvenile diabetes, or infant-dependent diabetes) is characterized by low insulin secretion, and patients with type 1 diabetes use insulin on a regular basis. There are no proven reasons or methods for avoiding

type 1 diabetes. Excessive urine excretion, thirst (polydipsia), constant starvation, loss of weight, changes in vision, and fatigue seem to be symptoms that we see in type 1 diabetes. The most prevalent form of diabetes is type 2 diabetes, which is known as insulin-dependent or adult-onset diabetes. For this form of diabetes, increased body weight and physical inactivity are largely responsible. The signs may be close to those of type 1 diabetes, but much less pronounced. Hyperglycemia with blood glucose value above normal but below diabetes diagnostic level is known as gestational diabetes. Diabetes is gestational during pregnancy. Women with gestational diabetes are increasingly vulnerable to complications during and after pregnancy. The risk of type 2 diabetes in the future also increases for those women and even their children who are suffering from gestational diabetes. Gestational diabetes is not diagnosed by the reported symptoms but by prenatal screening. Polyuria, polydipsia, fatigue, abrupt weight loss, polyphagia, vision blurring, genital thrush, swelling, delayed recovery, irritability, partial paresis, obesity, alopecia, and muscle stiffness are among the signs used in the data collection used in this research. The primary aims of this research are to predict diabetes at an early stage, so that people can take proper steps to control it, to find out the relation between different symptoms and factors that cause diabetes. Finally, this research will help us to ascertain the best machine learning classifier to predict diabetes.

SYSTEM ARCHITECTURE



METHODOLOGY

Data description:

The dataset consists of seventeen parameters, only one parameter is the response variable and the rest sixteen parameters are the predictor variables. Table 1 shows and describes briefly about these parameters.

Parameters Name	Parameters Type	Data Type	Possible Value
Age	Predictive	Integer	16-90
Polyuria	Predictive	Object	Yes, No
Sudden weight loss	Predictive	Object	Yes, No
Polydipsia	Predictive	Object	Yes, No
Weakness	Predictive	Object	Yes, No
Gender	Predictive	Object	Male, Female
Genital thrush	Predictive	Object	Yes, No
Polyphagia	Predictive	Object	Yes, No
Visual blurring	Predictive	Object	Yes, No
Partial paresis	Predictive	Object	Yes, No
Itching	Predictive	Object	Yes, No
Irritability	Predictive	Object	Yes, No
Muscle stiffness	Predictive	Object	Yes, No
Delayed healing	Predictive	Object	Yes, No
Obesity	Predictive	Object	Yes, No
Alopecia	Predictive	Object	Yes, No
Class	Responsive	Object	Positive, Negative

Table 1: Parameter Details

Classification Accuracy:

Classification accuracy is defined the numbers of the correct predictions are divided by the total number of inputs samples or total number of predictions made. It is mathematically given as:

$$CA = \frac{(NCP)}{(TNI)}$$

Allowing to this experiment the performance of artificial neural network is capable with highest accuracy is 98.9%. Decision tree is much closed result to ANN, having 92.7% accuracy.

Confusion Matrix:

Confusion matrix is doing well performance for the binary classification methods; we are also using binary classification in this research. Confusion matrix is giving the result as form of matrix, where describe the full performance of the proposed model. It gives us 4 values and two classes (actual class and predict class) in output. The mathematical representation of the average accuracy of confusion matrix shown below, where 'N' is the total number of inputs:

$$Accuracy = \frac{TP + FN}{N}$$

F1 Measure:

F1-score is also called F-measures. It is used for the measure of test's accuracy and identifying the number of true and positive of the precision and recall. It is the harmonic means value of the precision and recall. In this experiment the highest if values of AAN are 96%. Mathematically represent as:

$$FM = 2 \times \frac{precssion * recall}{precission - recall}$$

Precision:

Precision define is the fraction of true positive values among number of positive values predicted by the classifier. It is expressed as:

$$Precision = \frac{(TP)}{(TP) + (FP)}$$

Recall:

Recall, also referred to as sensitivity or true positive rate, and represents the ratio of correctly predicted positive outcomes to the total number of samples that are actually positive. Mathematically, it can be expressed as:

$$Precision = \frac{(TP)}{(TP) + (FN)}$$

TABLE III: RESULTS OF DIFFERENT CLASSIFIERS FOR PREDICTING DIABETES

Model	Accuracy	AUC	Label(No Diabetic=No, Diabetic=Yes)	Precision	Recall	F1-Score
LR	92%	0.97	No	0.89	0.87	0.88
			Yes	0.93	0.94	0.94
RF	98%	1.00	No	1.00	0.96	0.98
			Yes	0.98	1.00	0.99
DT	93%	0.96	No	0.93	0.89	0.91
			Yes	0.94	0.96	0.95
GP	88%	0.96	No	0.83	0.84	0.84
			Yes	0.92	0.91	0.91
AdaBoost	94%	0.98	No	0.93	0.89	0.91
			Yes	0.94	0.96	0.95
MLP	92%	0.98	No	0.84	0.96	0.90
			Yes	0.97	0.91	0.94
KNN	91%	0.97	No	0.82	0.93	0.87
			Yes	0.86	0.89	0.93
SVM	96%	0.99	No	0.95	0.93	0.94
			Yes	0.97	0.98	0.97
BNB	84%	0.94	No	0.74	0.87	0.80
			Yes	0.92	0.84	0.88
BC	92%	0.98	No	0.87	0.91	0.89
			Yes	0.95	0.93	0.94
QDA	95%	0.98	No	0.91	0.96	0.93
			Yes	0.98	0.95	0.96

Table 2: Results Of Different Classifiers for Predicting Diabetes

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 8, August : 2024

Service Provider:

In this module, the Service Provider has to login by using valid user name and password. After login successful he can do some operations such as Login, Browse Diabetic Data Sets and Train & Test, View Trained and Tested Accuracy in Bar Chart, View Trained and Tested Accuracy Results, View All Primary Stage Diabetic Prediction, Find Primary Stage Diabetic Prediction Type Ratio, View Primary Stage Diabetic Prediction Ratio Results, Download Predicted Data Sets, View All Remote Users.

View and Authorize Users:

In this module, the admin can view the list of users who all registered. In this, the admin can view the user's details such as, user name, email, address and admin authorize the users.

Remote User:

In this module, there are n numbers of users are present. User should register before doing any operations. Once user registers, their details will be stored to the database. After registration successful, he has to login by using authorized user name and password. Once Login is successful user will do some operations like REGISTER AND LOGIN, PREDICT PRIMARY STAGE DIABETIC STATUS, VIEW YOUR PROFILE.

RESULTS ANALYSIS

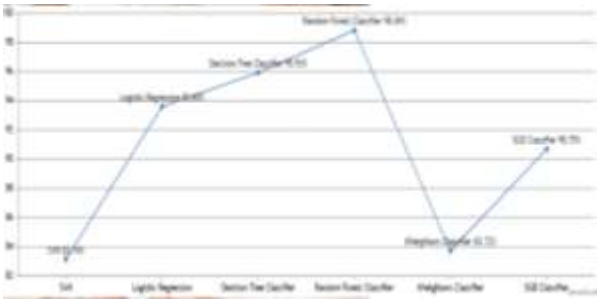
Prediction Type	Ratio
Positive	79.76878612716763
Negative	19.653179190751445

Prediction Ratio Details

TABLE II: CONFUSION MATRICES OF DIFFERENT CLASSIFIERS

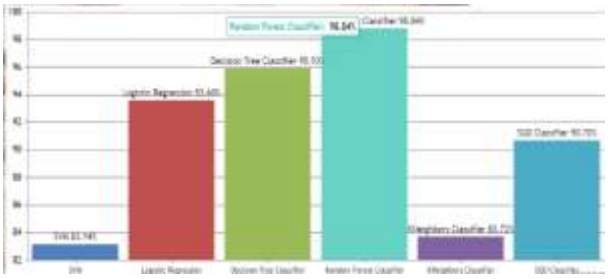
Model	Label	Predicted Negative	Predicted Positive
LR	Genuine Negative	39	6
	Genuine Positive	5	80
RF	Genuine Negative	43	2
	Genuine Positive	0	85
DT	Genuine Negative	40	5
	Genuine Positive	3	82
GP	Genuine Negative	38	7
	Genuine Positive	8	77
AdaBoost	Genuine Negative	40	5
	Genuine Positive	3	82
MLP	Genuine Negative	43	2
	Genuine Positive	8	77
KNN	Genuine Negative	42	3
	Genuine Positive	9	76
SVM	Genuine Negative	42	3
	Genuine Positive	2	83
BNB	Genuine Negative	39	6
	Genuine Positive	14	71
BC	Genuine Negative	41	4
	Genuine Positive	6	79
QDA	Genuine Negative	43	2
	Genuine Positive	4	81

Table 3: Confusion matrices of different classifiers

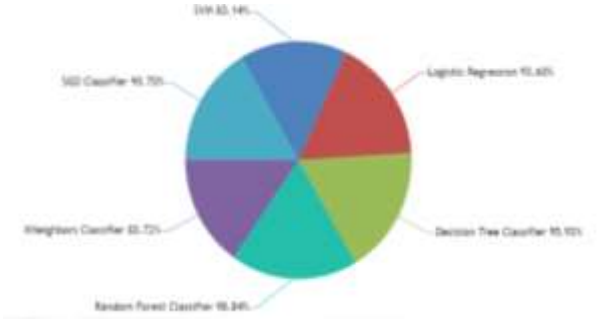


Line Chart Prediction Results

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 8, August : 2024



Bar Chart Prediction Results



Pie Chart Prediction Results

CONCLUSION

The early diagnosis of diabetes may play a significant role in the treatment process of diabetes. In this study, we have considered the early symptoms of diabetes as feature variables and employed machine learning algorithms to diagnose the presence of diabetes in a patient with these feature variables. Among the twelve classifiers that we have applied in this study, Random Forest has shown the best result in terms of different accuracy metrics. By diagnosing diabetes at an early stage, a patient can take necessary measures to control it, like taking low sugar foods, performing regular exercises, etc. In this study, we have considered the data of the patients of Sylhet Diabetes Hospital, Sylhet only. In the latter study, we shall accumulate data from a variety of regions of the country. Aggregating different classifiers using the Voting classifier sometimes enhances the performances of prediction. We can apply the Voting classifier to enhance the performance of predicting diabetic patients. Here, we have not filtered the most important features from the dataset to predict a diabetic patient. In the upcoming study, we can apply different feature selection techniques to find out the most relevant features for predicting diabetes.

FUTURE ENHANCEMENT

Expanding Data Scope: Broadening data collection beyond Sylhet Diabetes Hospital to encompass various regions will improve the model's generalizability to diverse populations.

Ensemble Learning: Utilizing a Voting Classifier, which combines predictions from multiple models like Random Forest, can potentially enhance overall accuracy and robustness.

Feature Selection: Implementing feature selection techniques can identify the most impactful factors for diabetes prediction, leading to a more streamlined model and potentially even better results.

Explainable AI: Integrating Explainable AI techniques can provide insights into the model's decision-making process, fostering trust and potentially leading to more targeted interventions for early diabetes management.

REFERENCES

1) Ziegler, A. G., & Nepom, G. T. (2010). Prediction and pathogenesis in type 1 diabetes. Immunity, 32(4), 468-478.
2) Wikipedia contributors. (2020, October 26). Diabetes. In Wikipedia, The Free Encyclopedia. Retrieved 06:20, October 31, 2020, from https://en.wikipedia.org/w/index.php?title=Diabetes&oldid=985520269
3) World Health Organization, (2020, October 28). Diabetes. Retrieved 06:20, October 31, 2020, from https://www.who.int/health-topics/diabetes
4) Singh, D. A. A. G., Leavline, E. J., & Baig, B. S. (2017). Diabetes prediction using medical data. Journal of Computational Intelligence in Bioinformatics, 10(1), 1-8.
5) Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. Computational and structural biotechnology journal, 15, 104-116.
6) El Jerjawi, N. S., & Abu-Naser, S. S. (2018). Diabetes prediction using artificial neural network.
7) Maniruzzaman, M., Rahman, M. J., Ahammed, B., Abedin, M. M. (2020). Classification and prediction of diabetes disease using machine learning paradigm. Health Information Science and Systems, 8(1), 7.
8) Sowjanya, K., Singhal, A., & Choudhary, C. (2015, June). MobDBTest: A machine learning based system for predicting diabetes risk using mobile devices. In 2015 IEEE International Advance Computing Conference (IACC) (pp. 397-402). IEEE.
9) Anand, R. S., Stey, P., Jain, S., Biron, D. R., Bhatt, H., Monteiro, K., ... & Chen, E. S. (2018). Predicting mortality in diabetic ICU patients using machine learning and severity indices. AMIA Summits on Translational Science Proceedings, 2018, 310.
10) Dagliati, A., Marini, S., Sacchi, L., Cogni, G., Teliti, M., Tibollo, V., ... & Bellazzi, R. (2018). Machine learning methods to predict diabetes complications. Journal of diabetes science and technology, 12(2), 295- 302.
11) Kumari, V. A., & Chitra, R. (2013). Classification of diabetes disease using support vector machine. International Journal of Engineering Research and Applications, 3(2), 1797-1801.
12) Plis, K., Bunescu, R., Marling, C., Shubrook, J., & Schwartz, F. (2014, June). A machine learning approach to predicting blood glucose levels for diabetes management. In Workshops at the Twenty-Eighth AAAI conference on artificial intelligence.
13) Kaur, H., & Kumari, V. (2020). Predictive modelling and analytics for

diabetes using a machine learning approach. Applied computing and informatics.

14) Islam, M. F., Ferdousi, R., Rahman, S., & Bushra, H. Y. (2020). Likelihood prediction of diabetes at early stage using data mining techniques. In Computer Vision and Machine Intelligence in Medical Image Analysis (pp. 113-125). Springer, Singapore.

15) Ahmed, T. M. (2016). Developing a predicted model for diabetes type 2 treatment plans by using data mining. Journal of Theoretical and Applied Information Technology, 90(2), 181.

16) Emon, M. U., Keya, M. S., Meghla, T. I., Rahman, M. M., Al Mamun, M. S., Kaiser, M. S. (2020, November). Performance Analysis of Machine Learning Approaches in Stroke Prediction. In 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA) (pp. 1464-1469). IEEE.

AUTHOR PROFILE

Mrs. Chepuri. Deepti, currently working as an Assistant Professor in the Department of Computer Science and Engineering, QIS College of Engineering and Technology, Ongole, Andhra Pradesh. She did her BTech from Uttar Pradesh Technical University, Lucknow, M.Tech from JNTUK, Kakinada. Her area of interest is Machine Learning, Artificial intelligence, Cloud Computing and Programming Languages.

Ms. Vutukuri Susmitha, currently pursuing Master of Computer Applications at QIS College of engineering and Technology (Autonomous), Ongole, Andhra Pradesh. She Completed B.Sc. in Statistics from Sri Harshini Degree College, Ongole, Andhra Pradesh. Her areas of interest are Machine learning & Cloud computing.