IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A COMPARISON OF DIABETIC PREDICTION USING ML APPROACHES

Shanthi[1] and Nancy Jasmine Goldena[2]

[1]PG Student, Department of Computer Applications and Research Centre, Sarah Tucker College (Autonomous), Tirunelveli, Tamil Nadu, India

[2]Associate Professor, Department of Computer Applications & Research Centre, Sarah Tucker College (Autonomous), Tirunelveli, Tamil Nadu, India

Abstract:

In India, diabetes affects approximately 30 million people around the world, and many more are at risk. In order to prevent diabetes and the health issues it is associated with, timely diagnosis and treatment is essential. This study tries to evaluate a person's risk of developing diabetes depends on their lifestyles and family history. In 2021, there are 7 million people died, according to the International Diabetes Federation (IDF). Through a change in lifestyle, nutrition, or medicine, early identification of patients with a high chance of developing T2D can lower the incidence of the illness. The diabetes dataset from the UCI machine learning repository is used in this analysis. C50 and GLM algorithms are used to do binary classification, and performance measures are used to assess the efficiency.

Keywords: Binary Classification, Comparison, Diabetes, Machine Learning, Supervised Learning

Introduction:

Diabetes is a metabolic disorder that raises blood sugar levels by either producing a large amount of insulin in the body. The most terrible disease on earth is diabetic. Therefore, early diagnosis of such a chronic metabolic disorder could assist medical professionals all over the world in preventing the death of individuals. Currently, with the rise of learning algorithms, intelligent systems, and neural systems, as well as their application in other fields, researchers have the chance to find a solution to this problem. In order to discover novel realities from current well-being-related formational indices, scientists use Machine Learning (ML) techniques and neural networks. This work may aid in the monitoring and detection of chronic conditions. Clinicians and patients may receive useful prediction data from the model regarding the probability to develop diabetes. A supervised learning system has been created to detect a person has diabetes or not using 952 occurrences, 18 distinct predictors, and the binary target.

Dataset Description:

The Diabetes dataset is taken from Kaggle machine learning repository. It has 952 occurrences with 18 unique predictive variables and one binary outcome. The dataset's goal is to predicting whether the patient is diabetic or not. The dataset has one dependent variable and a number of independent variables. The binary classification is going to be done in this investigation.

Outline of the Work:

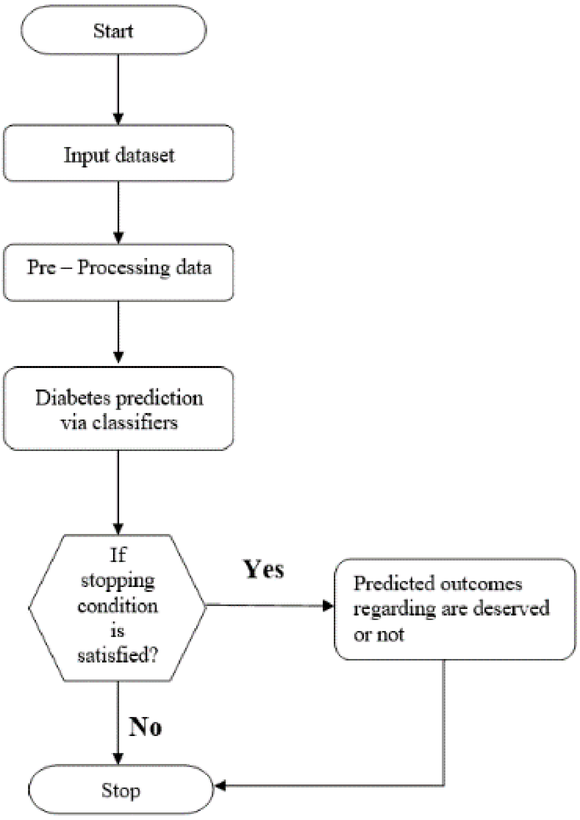
Start
Input dataset
Pre – Processing data
Diabetes prediction via classifiers
If stopping condition is satisfied?
Yes
Predicted outcomes regarding are deserved or not
No
Stop

Figure 1: Outline of the work

Feature Selection using Pearson Correlation:

The consistency between the independent parameters and their relationships is measured by the Pearson correlation coefficient. In order to determine how strong, the association between the two parameters is calculated by the correlation coefficient's value. If the coefficient is positive, the relationship between two variables is positive; if the coefficient is negative, the link between variables is negative.

$$\mathbf{r} = \frac{\mathbf{n}(\sum \mathbf{xy}) - (\sum \mathbf{x})(\sum \mathbf{y})}{\sqrt{[\mathbf{n}\sum \mathbf{x}^2 - (\sum \mathbf{x})^2][\mathbf{n}\sum \mathbf{y}^2 - (\sum \mathbf{y})^2]}}$$

Where,

r = Pearson Coefficient

n= number of the pairs of the stock

∑xy = sum of products of the paired stocks

∑x = sum of the x scores

∑y= sum of the y scores

∑x2 = sum of the squared x scores

∑y2 = sum of the squared y scores

Methodology:

Generalized Linear Model Classifier:

GLMs are very effective for fitting regression models, Regression models, as well as other complex models. Once a model has been fitted, the output value of a new data point can be predicted using the prediction function. The dependent variable's non-normal distribution is supported by the model. Through its extremely general model formulation, it covers a broad range of predictive methods, including threshold selection for binary data, loglinear models for data set, and linear regression for normally distributed responses. In this study GLM is used as a predictor for binary classification.

C 50 Classifier:

A decision tree is an especially special type of probability tree that enables users to choose an action to take. Although there are several decision tree solutions, the C50 technique is one of the most well-known. Because of C50 algorithm performs well for the majority of issue categories out of the box, it has become the industry standard for creating classification trees. The decision trees used in the C50 method typically perform nearly as well as more complex and sophisticated machine learning models (such as Neural Networks and Support Vector Machines), but they are considerably simpler to use and comprehend. In this study C 50 is used as a predictor for binary classification.

Performance Evaluation of Classifiers:

There are several methods for evaluating a model's efficiency, however one of the most popular is to simply compare the expected and actual outcomes. A model's accuracy is an indicator of how well it works. It represents the proportion of accurate predictions to all predictions. The method for accuracy is:

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Where,

TP = True Positive,

TN = True Negative,

FP = False Positive

FN = False Negative

Conclusion:

Finding the early signs of diabetes risk is one of the major medical disorders. This research aims to develop a framework that predicts the risk associated with type 2 diabetes. Two machine learning classification techniques were used in this study, and the outcomes were evaluated using several statistical metrics. The diabetes database also used the same algorithms. According to the experimental findings, Random Forest has the highest accuracy rate of all the models in our dataset (88.08%). For the diabetes dataset, random forest is likewise providing the maximum accuracy. All of the models achieved positive results for some parameters, such as precision, recall sensitivity, etc., using two distinct machine learning techniques.

Future Enhancement:

Other ML methods may be used in the future. This data set just includes two classes: if the patient has diabetes or not, but additional multiclass classification can be done, like diabetes stages. Future feature selection and classification techniques can be used.

References:

1. http://diabetesindia.com
2. https://my.clevelandclinic.org/health/diseases/7104-diabetes-mellitus-an-overview
3. https://www.diabetes.co.uk/diabetes_care/blood-sugar-level-ranges.html
4. Deberneh, H.M.; Kim, I.; Park, J.H.; Cha, E.; Joung, K.H.; Lee, J.S.; Lim, D.S. 1233-P: Prediction of type 2 diabetes occurrence using machine learning model. Am. Diabetes Assoc. 2020, 69, 1233.
5. Buch, V.; Varughese, G.; Maruthappu, M. Artificial intelligence in diabetes care. Diabet. Med. 2018, 35, 495–497.
6. Dankwa-Mullan, I.; Rivo, M.; Sepulveda, M.; Park, Y.; Snowdon, J.; Rhee, K. Transforming diabetes care through artificial intelligence: The future is here. Popul. Health Manag. 2019, 22, 229–242.
7. Woldaregay, A.Z.; Årsand, E.; Botsis, T.; Albers, D.; Mamykina, L.; Hartvigsen, G. Data-driven blood glucose pattern classification and anomalies detection: Machine-learning applications in type 1 diabetes. J. Med. Internet Res. 2019, 21, e11030. Int. J. Environ. Res. Public Health 2021, 18, 3317 13 of 14
8. Maniruzzaman Kumar, N.; Abedin, M.; Islam, S.; Suri, H.S.; El-Baz, A.S.; Suri, J.S. Comparative approaches for classification of diabetes mellitus data: Machine learning paradigm. Comput. Methods Programs Biomed. 2017, 152, 23–34.
9. Kavakiotis, I.; Tsave, O.; Salifoglou, A.; Maglaveras, N.; Vlahavas, I.; Chouvarda, I. Machine learning and data mining methods in diabetes research. Comput. Struct. Biotechnol. J. 2017, 15, 104–116.
10. https://www.kaggle.com/code/mirichoi0218/classification-diabetes-or-not-with-basic-12ml

11. Swapna, G., Vinayakumar R., Soman K. P. (2018) "Diabetes detection using deep learning algorithms." ICT Express 4 (4): 243-246.
12. Perveen, S., Shahbaz, M., Keshavjee, K., Guergachi, A. (2019) "Metabolic Syndrome and Development of Diabetes Mellitus: Predictive Modeling Based on Machine Learning Techniques." IEEE Access 7: 1365-1375.
13. Deberneh, H.M.; Kim, I.; Park, J.H.; Cha, E.; Joung, K.H.; Lee, J.S.; Lim, D.S. 1233-P: Prediction of type 2 diabetes occurrence using machine learning model. Am. Diabetes Assoc. 2020, 69, 1233.
14. Liu, X.; Faes, L.; Kale, A.U.; Wagner, S.K.; Fu, D.J.; Bruynseels, A.; Mahendiran, T.; Moraes, G.; Shamdas, M.; Kern, C.; et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: A systematic review and meta-analysis. Lancet Digit. Health 2019, 1, e271–e297.
15. Hutchinson, M.S.; Joakimsen, R.M.; Njølstad, I.; Schirmer, H.; Figenschau, Y.; Svartberg, J.; Jorde, R. Effects of age and sex on estimated diabetes prevalence using different diagnostic criteria: The Tromsø OGTT Study. Int. J. Endocrinol. 2013, 2013, 1–9.