Diabetes Prediction Using an Assemblage of ML Classifiers

Neha Titarmare[1], Shreya Khewale[2], Srushti Jamnerkar[3], Shiba Malkhede[4], Anushka Meshram[5], Pranjali Urkude[6]

[1,2,3,4,5,6]*Computer science and Engineering, Rajiv Gandhi college of engineering and Research, Wanadongri, Nagpur, India*

Abstract— Many different machine learning techniques are used in a range of disciplines to do predictive analytics over massive data. Despite the challenges, predictive analytics in healthcare can ultimately help practitioners make quick decisions regarding the health and treatment of patients based on enormous amounts of data. Six different machine learning algorithms are used in this research effort, which addresses predictive analytics in healthcare. A collection of patient medical records is acquired for experimental purposes, and six different machine-learning algorithms are applied to the dataset. It is discussed and contrasted how well the used algorithms perform and how accurate they are. The algorithm that is most effective for diabetes prediction can be determined by comparing the various machine learning methods employed in this study. Through the use of machine learning techniques, this initiative seeks to aid medical professionals in the early diagnosis and treatment of diabetes.

Keywords- diabetes, machine learning algorithms, healthcare

I.INTRODUCTION

Machine learning:

We were able to create a machine learning model (random forest, the best one) using all the patient records, which accurately predicts if the dataset's patients have diabetes. Using data analysis and visualisation, we were also able to glean some conclusions from the information.

ALGORITHMS:

Naive Bayes (NB): Undoubtedly one of the simplest and most effective classification algorithms is the Naive Bayes Classifier. It makes it easier to create effective machine learning models that can make accurate predictions. It provides predictions based on the likelihood that an object will occur because it is a probabilistic classifier.

SVM: SVMs can do classification and regression on both linear and non-linear data in machine learning. Web sites, intrusion detection, face recognition, email classification, gene classification, and handwriting recognition are just a few of the applications that use them.

LR: Regular Regression Logistic regression training is very efficient and simpler to implement and analyse. It swiftly classifies records that are unfamiliar. When the dataset can be partitioned linearly, it functions well. A trait's significance can be determined using the model coefficients.

Random Forest : A machine learning technique called a random forest is used to solve classification and regression problems. Ensemble learning is used here, a method for combining various classifiers to tackle challenging tasks. The decision trees that can be used in a random forest algorithm are numerous.

K Nearest Neighbor (KNN): A simple technique for handling classification and regression problems is the supervised k-nearest neighbours (KNN) algorithm.

Deterministic Tree (DT): Interpretation is aided by deterministic trees. It is straightforward enough for a layperson to understand and may be illustrated graphically. It also represents the criteria that physicians employ to make medical diagnoses. The greedy approach used in recursive binary splitting ensures that the best split is chosen at each level of tree construction rather than one that will produce a better tree later on. Patients can therefore undertake laboratory tests in the order of the nodes and can wrap up the process earlier if certain criteria are satisfied.

II.LITREATURE SURVEY

“A Comprehensive Review of Various Diabetic Prediction Models”

The 102538 values and 49 features in the dataset [1]Negi and Jaiswal's [21] creation. Of the people in this sample, 64419 had diabetes; the rest did not. Missing values were replaced, and nominal data was transformed to numerical data, using preprocessing procedures. The wrapper feature selection methodology and the ranker method were used to select the pertinent features from the dataset. An further use of a small ensemble of classifiers produced an accuracy of 72%.

[2] On a locally accessible dataset that was obtained from a hospital in Germany, Malik et al. used decision trees, k-nearest neighbour, and random forest. In their research, Soltani and Jafarian used the probabilistic neural network.

[3**]** Qawqzeh et al. employed the logistic regression classification technique to categorise the diabetes data. While training data has 459 patients, testing data has 128 cases. The authors' classification accuracy using logistic regression was 92%. The main problem with the model was that it couldn't be validated because it wasn't compared to other diabetic prediction models.

[4] Tafa et al. divided the dataset in half for training and half for testing. The model was released using a combination of naive Bayes and support vector machine techniques for the purpose of predicting diabetes. A dataset that was compiled from three distinct sources was used to evaluate the proposed model. Eight characteristics and 402 people were included in the dataset, 80 of whom had type 2 diabetes. When applied to the dataset, the accuracy of the combined Nave Bayes and SVM approach was 97.6%, which is much greater than the accuracy of the separate algorithms' results, which were Nave Bayes' 94.52 and SVM's 95.52%. The authors have not included a description of a preprocessing technique to remove undesired values from the dataset.

[5]Karan et al. demonstrated a novel method for the diagnosis of diabetes by developing a distributed end-to-end three-level unavoidable healthcare system architecture utilising artificial neural network (ANN) computation. The most basic level of vital sign monitoring is done with sensors and wearable technology. PDAs and PCs are examples of client-side devices that serve as a mediator and conduit for communication between the primary and final levels at level 2. The third level end, which provides customers with social welfare administrations and database operations, includes powerful desktop servers. Techniques based on artificial neural networks are utilised to detect diseases in both the advanced and advanced phases. Artificial neural network computations underpin the client-server concept. This approach enhances computations and system communications on both the user and server sides by depending on the notion of illnesses.

[6]Sisodia used Nave Bayes, decision trees, and support vector machine learning methods on the Pima Indians Diabetes Dataset. The best classifier for predicting diabetes was Nave Bayes. The dataset was divided into 10 equal sections, nine of which were used for training and the tenth for testing. Sisodia used a tenfold cross-validation technique. Several evaluation metrics, including accuracy, precision, recall, and area under the curve, were applied to the diabetes prediction process.

[7]Hussain and Naaz reviewed several machine learning algorithms and evaluated the efficacy of random forest, naive bayes, and neural networks. The Matthews correlation coefficient was employed by the authors to assess various machine learning techniques.

[8] Naive Bayes, random forest, and logistic regression were applied to the Pima Indians Diabetes Dataset by Kumari et al. They then contrasted these three approaches with an ensemble approach, and found that the ensemble approach outperformed the individual approaches with an accuracy of 79%.

[9] Deep learning, or a neural network—a multilayer, feed-forward network—was used by Olaniyi and Adnan. The Pima Indians Diabetes Dataset was utilised by the authors to test and train the algorithm. The dataset was separated into 500 values for training and 268 values for testing. In order to attain numerical stability, the dataset had to be normalised before any preprocessing operations could be carried out. By dividing each attribute by its associated amplitude, the dataset values were all converted to ranges between 0 and 1, in order to achieve dataset normalisation. The prediction rate for the writers was 82% accurate.

[10] The diabetes dataset was categorised by Gupta et al. using Naive Bayes and SVM. The support vector machine classifier outperformed the Nave Bayes method after employing both classification algorithms, according to the authors' K-fold cross-validation model, which was used for both training and testing.

III.PROBLEM STATEMENT

To employ machine learning, there needs be a vast amount of data. The data is generally scarce and depends on the illness. There are also significantly more samples without illnesses than there are with real diseases.

IV.PROPOSED WORK

Proposed Approach/Work

The paper's goal is to investigate models that can predict diabetes more precisely. We experimented with various classification and ensemble approaches to predict diabetes. The parts that follow provide a quick overview of the time period.

Pima Indian Diabetes Dataset, a UCI repository, is where the data for this dataset was collected. The collection contains a lot of information about 768 patients.

Sr.no.	Attributes
1.	Pregnancy
2.	Glucose
3.	Blood pressure
4.	Skin thickness
5.	Insulin
6	BMI
7.	Diabetes Pedigree Function
8.	Age

Table 1: Dataset Description

The ninth characteristic for each data point is the class variable. This class variable displays the result for diabetics (0 or 1), indicating whether it is positive or negative for diabetes.

Diabetes sufferers' distribution Although the dataset was significantly skewed, with 268 classes labelled as positive for diabetes and 268 classes labelled as negative for no diabetes, we nevertheless managed to develop a model that predict diabetes.

Proposed Architecture

Data Preparation: The most crucial step is data preprocessing. The majority of healthcare data has missing information and other impurities that can limit its usefulness. Data preparation is done to enhance the quality and effectiveness of the results produced through mining. When applying machine learning techniques to the dataset, For sound prediction and accurate findings, the strategy is crucial. The Pima Indian diabetes dataset requires two types of pre-processing.

Remove all instances where the value is zero (0) using the Missing Values removal function. Since zero is not a valid value, this instance is ended. The process of selecting a feature subset, which minimises the dimensionality of the data and helps to work more quickly, is created by removing unimportant features.

Data splitting: Data is standardised for use in both training and testing the model after data cleaning. We divide the data into training and test sets and then train the algorithm using the training set only. Based on logic, algorithms, and feature values in the training data, this process will create the training model. The fundamental goal of normalisation is to scale all qualities equally.

Apply machine learning: The machine learning technique is applied after the data is prepared. To predict diabetes, we employ a variety of classification and ensemble algorithms. Application of machine learning techniques is primarily intended to assess their efficiency and accuracy, and identify key features that are crucial for accurate prediction.

V.AIM AND OBJECTIVE

AIM :

Utilizing machine learning techniques, this initiative seeks to assist medical professionals in the early diagnosis and treatment of diabetes.

OBJECTIVE :

Three distinct machine learning algorithms are applied to a dataset of patient medical records to produce the dataset. The effectiveness and precision of the used algorithms are explored and contrasted. This study compares various machine learning algorithms to see which algorithm is most effective in predicting diabetes and determining whether a patient has the disease or not. Care must be taken when handling the vast and sensitive data that the healthcare business generates. One of the numerous fatal diseases that are spreading around the globe is diabetes mellitus. Medical professionals want a reliable diabetes prediction system. The data can be analysed using a variety of machine learning techniques, which can then be combined to produce useful insights. If specific data mining techniques are applied to it, the accessibility and availability of vast volumes of data will be able to give us meaningful knowledge. The main goal is to find new patterns, comprehend these patterns, and then give users relevant information that is important to them. Diabetes is a risk factor for blindness, kidney impairment, nerve damage, and heart disease. Effective data mining for diabetes is a critical issue. The proper procedures and methods for the accurate classification of the Diabetes dataset and the extraction of useful patterns will be found using data mining techniques and processes. In this study, diabetes was predicted using medical bioinformatics methods. Using mining techniques and the WEKA programme, diabetes was identified. The Pima Indian diabetes database was obtained from the UCI repository and analysed after that. The dataset was examined and assessed in order to build a potent model that forecasts and recognises diabetic illness. Using an approach similar to bootstrapping, we aim to increase accuracy in this study, after which we compare the accuracy of Naive Bayes, Decision Trees, and (KNN) models.

VI.CONCLUSION

Diabetes is a condition that affects adults quite frequently today. It is crucial that this illness be discovered as soon as possible. The creation of the best algorithm and prediction for diabetic patients is the main objective of this research project. Over the previous five years, the accuracy of machine learning techniques was examined. The authors have therefore proposed a soft voting classifier model employing an ensemble of three machine learning algorithms, including Naive Bayes, Logistic Regression, and Random Forest. Before the suggested model was applied to the breast cancer dataset, the experimental dataset for Pima Indians with diabetes was employed. On the Pima Indians diabetes dataset, the ensemble soft voting classifier provided findings that were 79.08% accurate, while on the breast cancer dataset, it provided results that were 97.02% accurate. This accuracy could be improved in the future by utilising various deep learning models.

VII.ACKNOWLEDGEMENT

We are grateful beyond measure to our mentor Prof. Neha Titarmare, under whose direction our project evolved from a simple concept into reality. We are appreciative of our guide's invaluable advice and never-ending support in illuminating us. We express our gratitude to our guide for giving us a lot of help and valuable time. We owe our leader a debt of appreciation for always inspiring us to accomplish our goals.

Dr. Hemant Turkar, HOD CSE, RGCER, has our sincere gratitude for his gracious assistance and promptness.

We would like to thank Dr. Sachin Mahakalkar, Dean Academics at RGCER, for his unwavering support, forethought, and inspiration.

We would like to thank Dr. Manali M. Kshirsagar, Principal RGCER, for her unwavering encouragement and assistance.

Finally, we would like to express our gratitude to everyone who contributed to the success of our project, whether they were directly or indirectly involved.

REFERENCE

[1] A. Belle, R. Thiagarajan, S. M. R. Soroushmehr, F. Navidi, D. A. Beard, and K. Najarian, “Big Data Analytics in Healthcare,” Hindawi Publ. Corp., vol. 2015, pp. 1–16, 2015.

[2] J. Andreu-Perez, C. C. Y. Poon, R. D. Merrifield, S. T. C. Wong, and G.-Z. Yang, “Big Data for Health,” IEEE J. Biomed. Heal. Informatics, vol. 19, no. 4, pp. 1193–1208, 2015

[3] E. Ahmed et al., “The role of big data analytics in Internet of Things,” Comput. Networks, vol. 129, no. December, pp. 459–471, 2017

[4] “The big-data revolution in US health care: Accelerating value and innovation | McKinsey & Company.” [Online]. Available: https://www. mckinsey.com/industries/healthcare-systems-andservices/ our-insights/the-big-data-revolution-in-us-health-care. [Accessed: 12-May-2018]..

[5] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, “Disease Prediction by Machine Learning over Big Data from Healthcare Communities,” IEEE Access, vol. 5, no. c, pp. 8869–8879, 2017.

[6] L. Zhou, S. Pan, J. Wang, and A. V. Vasilakos, “Machine learning on big data: Opportunities and challenges,” Neurocomputing, vol. 237, pp. 350–361, May 2017.

[7] J. B. Heaton, N. G. Polson, and J. H. Witte, “Deep learning for finance: deep portfolios,” Appl. Stoch. Model. Bus. Ind., vol. 33, no. 1, pp. 3–12, Jan. 2017.

[8] K. Lin, M. Chen, J. Deng, M. M. Hassan, and G. Fortino, “Enhanced Fingerprinting and Trajectory Prediction for IoT Localization in Smart Buildings,” IEEE Trans. Autom. Sci. Eng., vol. 13, no. 3, pp. 1294–1307, Jul. 2016

[9] K. Lin, J. Luo, L. Hu, M. S. Hossain, and A. Ghoneim, “Localization Based on Social Big Data Analysis in the Vehicular Networks,” IEEE Trans. Ind. Informatics, vol. 13, no. 4, pp. 1932– 1940, Aug. 2017.

[10] P. A. Chiarelli, J. S. Hauptman, and S. R. Browd, “Machine Learning and the Prediction of Hydrocephalus,” JAMA Pediatr., vol. 172, no. 2, p. 116, Feb. 2018.

[11] A. Jindal, A. Dua, N. Kumar, A. K. Das, A. V. Vasilakos, and J. J. P. C. Rodrigues, “Providing Healthcare-as-a-Service Using Fuzzy Rule-Based Big Data Analytics in Cloud Computing,” IEEE J. Biomed. Heal. Informatics, pp. 1–1, 2018.

[12] N. M. S. kumar, T. Eswari, P. Sampath, and S. Lavanya, “Predictive Methodology for Diabetic Data Analysis in Big Data,” Comput. Sci., vol. 50, pp. 203–208, Jan. 2015.

[13] J. Zheng and A. Dagnino, “An initial study of predictive machine learning analytics on large volumes of historical data for power system applications,” in 2014 IEEE International Conference on Big Data (Big Data), 2014, pp. 952–959.

[14] International Journal of Advanced Computer and Mathematical Sciences. Bi Publication-BioIT Journals, 2010.

[15] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, “Disease Prediction by Machine Learning Over Big Data From Healthcare Communities,” IEEE Access, vol. 5, pp. 8869–8879, 2017.