International Journal of Engineering Research and Science & Technology

IJERST

ISSN : 2319-5991

www.ijerst.com

Email: editor@ijerst.com or editor.ijerst@gmail.com

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 17, Issue 2, 2024

DIABETES DISEASE PREDICTION USING MACHINE LEARNING ALGORITHMS

[1] Mr.K. LAKSHMANA REDDY, [2] YERRAMSETTI ANITHA

[1]Associate Professor, S.V.K.P & Dr K.S. Raju Arts & Science College(A), Penugonda, W.G.District, Andhra Pradesh, klreddyin@gmail.com

[2]PG, scholar, S.V.K.P & Dr K.S. Raju Arts & Science College(A) Penugonda, W.G.District, Andhra Pradesh, yerramsettianitha@gmail.com

ABSTRACT

This paper deals with the prediction of Diabetes Disease by performing an analysis of five supervised machine learning algorithms, i.e. K-Nearest Neighbors, Naive Baye, Decision Tree Classifier, Random Forest and Support Vector Machine. Further, by incorporating all the present risk factors of the dataset, we have observed a stable accuracy after classifying and performing cross-validation. We managed to achieve a stable and highest accuracy of 76% with KNN classifier and remaining all other classifiers also give a stable accuracy of above 70%. We analyzed why specific Machine Learning classifiers do not yield stable and good accuracy by visualizing the training and testing accuracy and examining model overfitting and model underfitting. The main goal of this paper is to find the most optimal results in terms of accuracy and computational time for Diabetes disease prediction.

1.INTRODUCTION

In this day and age, one of the most notorious diseases to have taken the world by storm is Diabetes, which is a disease which causes an increase in blood glucose levels as a result of the absence or low levels of insulin. Due to the many criterions to be taken into consideration for an individual to harbour this disease, it's detection and prediction might be tedious or sometimes inconclusive. Nevertheless, it isn't impossible to detect it, even at an early stage. Federation- IDF). 79% of the adult population were living in the countries with the low and middle-income groups. It is estimated that by the year 2045 approx. 700 million people will have diabetes (IDF).

Diabetes is increasing day by day in the world because of environmental,

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 17, Issue 2, 2024

genetic factors. The numbers are rising rapidly due to several factors which includes unhealthy foods, physical inactivity and many more. Diabetes is a hormonal disorder in which the inability of the body to produce insulin causes the metabolism of sugar in the body to be abnormal, thereby, raising the blood glucose levels in the body of a particular individual. Intense hunger, thirst and frequent urination are some of the observable characteristics. Certain risk factors such as age, BMI, Glucose Levels, Blood Pressure, etc., play an important role to the contribution of the disease.

In the Fig. 1 we can see that the number of cases is rising every year and there is not slowing down in the active cases. It is a very crucial thing to worry as diabetes has become one of the most dangerous and fastest diseases to take the lives of many individuals around the globe.

Machine Learning is very popular these days as it is used everywhere, where a large amount of data is present, and we need some knowledge from it. Generally, we can categorize the Machine Learning algorithms in two types but not limited to-

• Unsupervised Learning: In unsupervised learning, the information is not labelled and also not trained. Here, we just put the data in action to find some patterns if possible.

• Supervised Learning: In supervised learning, we train the model based on the labels attached to the information and based on that we classify or test the new data with labels.

With the rise of Machine Learning and its relative algorithms, it has come to light that the significant problems and hindrances in its detection faced earlier, can now be eased with much simplicity, yet, giving a detailed and accurate outcome. As of the modern-day, it is comprehended that Machine Learning has become even more effective and helpful in collaboration with the domain of Medicine. Early determination of a disease can be made possible through machine learning by studying the characteristics of an individual. Such early tries can lead to the inhibition of disease as well as obstruction of permitting the disease to reach a critical degree. The work which will be described in this paper is to perform the diabetes disease prediction using machine learning algorithms for early care of an individual.

2.LITERATURE SURVEY

Decision tree classifiers

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 17, Issue 2, 2024

Decision tree classifiers are used successfully in many diverse areas. Their most important feature is the capability of capturing descriptive decision making knowledge from the supplied data. Decision tree can be generated from training sets. The procedure for such generation based on the set of objects (S), each belonging to one of the classes C1, C2, …, Ck is as follows:

Step 1. If all the objects in S belong to the same class, for example Ci, the decision tree for S consists of a leaf labeled with this class

Step 2. Otherwise, let T be some test with possible outcomes O1, O2,…, On. Each object in S has one outcome for T so the test partitions S into subsets S1, S2,… Sn where each object in Si has outcome Oi for T. T becomes the root of the decision tree and for each outcome Oi we build a subsidiary decision tree by invoking the same procedure recursively on the set Si.

Gradient boosting

Gradient boosting is a machine learning technique used in regression and classification tasks, among others. It gives a prediction model in the form of an ensemble of weak prediction models, which are typically decision trees.[1][2] When a decision tree is the weak learner, the resulting algorithm is called gradient-boosted trees; it usually outperforms random forest. A gradient-boosted trees model is built in a stage-wise fashion as in other boosting methods, but it generalizes the other methods by allowing optimization of an arbitrary differentiable loss function.

K-Nearest Neighbors (KNN)

- Simple, but a very powerful classification algorithm
- Classifies based on a similarity measure
- Non-parametric
- Lazy learning
- Does not “learn” until the test example is given
- Whenever we have a new data to classify, we find its K-nearest neighbors from the training data

Example

- Training dataset consists of k-closest examples in feature space
- Feature space means, space with categorization variables (non-metric variables)
- Learning based on instances, and thus also works lazily because instance close to the input vector for

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 17, Issue 2, 2024

test or prediction may take time to occur in the training dataset

Logistic regression Classifiers

Logistic regression analysis studies the association between a categorical dependent variable and a set of independent (explanatory) variables. The name *logistic regression* is used when the dependent variable has only two values, such as 0 and 1 or Yes and No. The name *multinomial logistic regression* is usually reserved for the case when the dependent variable has three or more unique values, such as Married, Single, Divorced, or Widowed. Although the type of data used for the dependent variable is different from that of multiple regression, the practical use of the procedure is similar.

Logistic regression competes with discriminant analysis as a method for analyzing categorical-response variables. Many statisticians feel that logistic regression is more versatile and better suited for modeling most situations than is discriminant analysis. This is because logistic regression does not assume that the independent variables are normally distributed, as discriminant analysis does.

This program computes binary logistic regression and multinomial logistic regression on both numeric and categorical independent variables. It reports on the regression equation as well as the goodness of fit, odds ratios, confidence limits, likelihood, and deviance. It performs a comprehensive residual analysis including diagnostic residual reports and plots. It can perform an independent variable subset selection search, looking for the best regression model with the fewest independent variables. It provides confidence intervals on predicted values and provides ROC curves to help determine the best cutoff point for classification. It allows you to validate your results by automatically classifying rows that are not used during the analysis.

Naive Bayes

The naive bayes approach is a supervised learning method which is based on a simplistic hypothesis: it assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature .

Yet, despite this, it appears robust and efficient. Its performance is comparable to other supervised learning techniques. Various reasons have been advanced in the literature. In this tutorial, we highlight an

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 17, Issue 2, 2024

explanation based on the representation bias. The naive bayes classifier is a linear classifier, as well as linear discriminant analysis, logistic regression or linear SVM (support vector machine). The difference lies on the method of estimating the parameters of the classifier (the learning bias).

While the Naive Bayes classifier is widely used in the research world, it is not widespread among practitioners which want to obtain usable results. On the one hand, the researchers found especially it is very easy to program and implement it, its parameters are easy to estimate, learning is very fast even on very large databases, its accuracy is reasonably good in comparison to the other approaches. On the other hand, the final users do not obtain a model easy to interpret and deploy, they does not understand the interest of such a technique.

Thus, we introduce in a new presentation of the results of the learning process. The classifier is easier to understand, and its deployment is also made easier. In the first part of this tutorial, we present some theoretical aspects of the naive bayes classifier. Then, we implement the approach on a dataset with Tanagra. We compare the obtained results (the parameters of the model) to those obtained with other linear approaches such as the logistic regression, the linear discriminant analysis and the linear SVM. We note that the results are highly consistent. This largely explains the good performance of the method in comparison to others. In the second part, we use various tools on the same dataset (**Weka 3.6.0**, **R 2.9.2**, **Knime 2.1.1**, **Orange 2.0b** and **RapidMiner 4.6.0)**. We try above all to understand the obtained results.

Random Forest

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the mean or average prediction of the individual trees is returned. Random decision forests correct for decision trees' habit of overfitting to their training set. Random forests generally outperform decision trees, but their accuracy is lower than gradient boosted trees. However, data characteristics can affect their performance.

The first algorithm for random decision forests was created in 1995 by Tin Kam

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 17, Issue 2, 2024

Ho[1] using the random subspace method, which, in Ho's formulation, is a way to implement the "stochastic discrimination" approach to classification proposed by Eugene Kleinberg.

An extension of the algorithm was developed by Leo Breiman and Adele Cutler, who registered "Random Forests" as a trademark in 2006 (as of 2019, owned by Minitab, Inc.).The extension combines Breiman's "bagging" idea and random selection of features, introduced first by Ho[1] and later independently by Amit and Geman[13] in order to construct a collection of decision trees with controlled variance.

Random forests are frequently used as "blackbox" models in businesses, as they generate reasonable predictions across a wide range of data while requiring little configuration.

SVM

In classification tasks a discriminant machine learning technique aims at finding, based on an *independent and identically distributed* (*iid*) training dataset, a discriminant function that can correctly predict labels for newly acquired instances. Unlike generative machine learning approaches, which require computations of conditional probability distributions, a discriminant classification function takes a data point x and assigns it to one of the different classes that are a part of the classification task. Less powerful than generative approaches, which are mostly used when prediction involves outlier detection, discriminant approaches require fewer computational resources and less training data, especially for a multidimensional feature space and when only posterior probabilities are needed. From a geometric perspective, learning a classifier is equivalent to finding the equation for a multidimensional surface that best separates the different classes in the feature space.

SVM is a discriminant technique, and, because it solves the convex optimization problem analytically, it always returns the same optimal hyperplane parameter—in contrast to *genetic algorithms* (*GAs*) or *perceptrons*, both of which are widely used for classification in machine learning. For perceptrons, solutions are highly dependent on the initialization and termination criteria. For a specific kernel that transforms the data from the input space to the feature space, training returns uniquely defined SVM model parameters for a given training set, whereas the

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 17, Issue 2, 2024

perceptron and GA classifier models are different each time training is initialized. The aim of GAs and perceptrons is only to minimize error during training, which will translate into several hyperplanes' meeting this requirement.

3.EXISTING SYSTEM

In [2], they have used the WEKA tool for data analytics for diabetes disease prediction on Big Data of healthcare. They used the publicly available dataset from UCI and applied different machine learning classifiers on it. The classifiers which they incorporated are Naive Bayes, Support Vector Machine, Random Forest and Simple CART.

Their approach starts with accessing the dataset, preprocess it in Weka tool and then did the 70:30 train and test split for applying different machine algorithms. They did not go with the cross-validation step as it is imperative to get the optimal and accurate results as well.

The authors in [3], also used the publicly available dataset named as Pima Indians Diabetes Database for performing their experiment. Their framework of performing the prediction starts with the dataset selection and then with data pre-processing. Once the data was preprocessed, they applied three classification algorithms, i.e. naive Bayes, SVM and Decision tree. As they incorporated different evaluation metrics, they did compare the different performance measure and comparatively analyzed the accuracy. The highest accuracy achieved with their experiment was 76.30%. Like [2] they have also not practised Cross-validation.

In [4], the authors proposed the neural network-based diabetes disease prediction on Indians Pima Diabetes Dataset. They have used several hidden layers to find patterns in the data, and with the help of those patterns, they predicted the outcome. They name their proposed algorithms as ADAP, which is a custom neural network with multiple partitions and with the set of association weights and units. They managed to achieve a crossover point for sensitivity, and specificity at 0.76 and are trying to precise their result in future.

Advantages

1). The system more effective due to fitting datasets for different ML Models by Applying Machine Learning Algorithms.

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 17, Issue 2, 2024

2). The Early determination of a disease can be made possible through machine learning by studying the characteristics of an individual in the proposed system.

4.PROPOSED SYSTEM

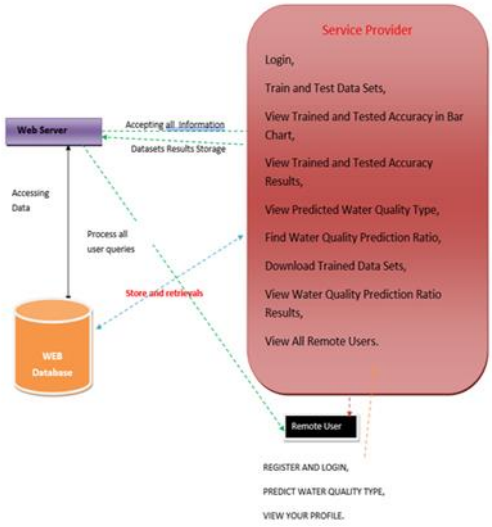
To perform our experiment, we have used a publicly available dataset named as Pima Indians Diabetes Database [4]. This dataset includes a various diagnostic measure of diabetes disease. The dataset was originally from the National Institute of Diabetes and Digestive and Kidney Diseases. All the recorded instances are of the patients whose age are above 21 years old. Our proposed model exists of 5 phases which are shown in the proposed system.

Disadvantages

1). There are no techniques and models for analyzing large scale datasets in the existing system.

2). There is no facility for diabetes dataset in collaboration with a hospital or a medical institute and will try to achieve better results.

5.ARCHITECTURE DIAGRAM



6.MODULES

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
Train and Test Data Sets,View Trained and Tested Accuracy in Bar Chart, View Trained and Tested Accuracy Results, Find Diabetic Status From Data Set Details, Find Diabetic Ratio on Data Sets, View All Emergency For Diabetic Treatment, Download Trained Data Sets, View Diabetic Ratio Results, View All Remote Users.

View and Authorize Users

In this module, the admin can view the list of users who all registered.

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 17, Issue 2, 2024

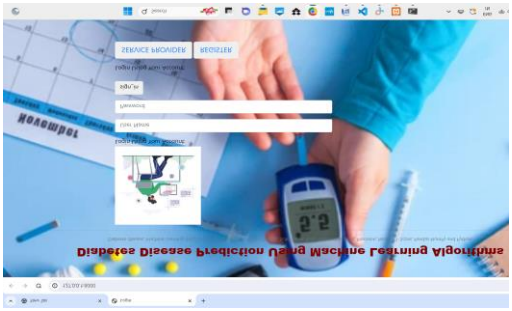
In this, the admin can view the user's details such as, user name, email, address and admin authorizes the users.

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 POST DIABETIC DATA SETS,SEARCH AND PREDICT DIABETIC STATUS, VIEW YOUR PROFILE.

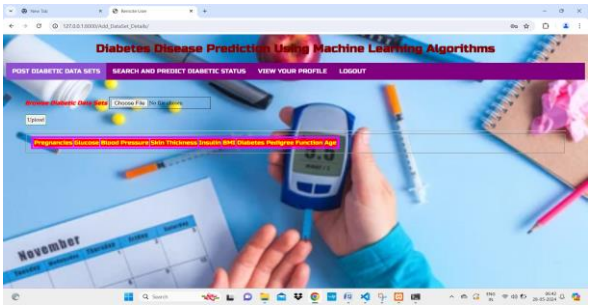
7.OUTPUT SCREENS:

Login Screen:

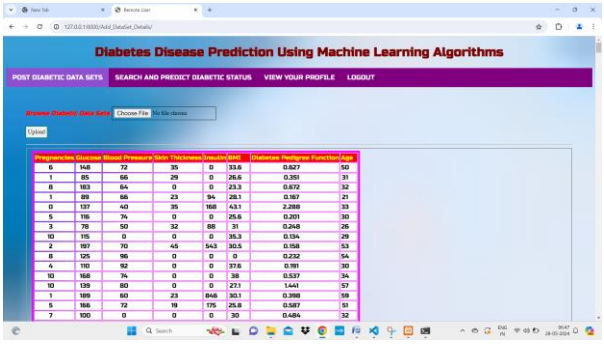


Register Screen:

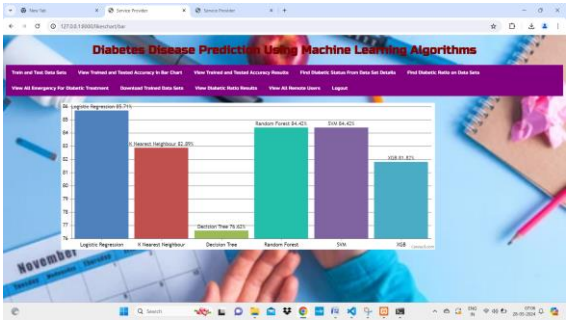
ADMIN HOME PAGE



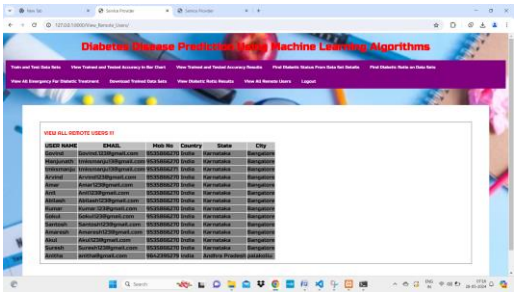
Train & Test Data Screen:



Accuracy Bar Chart Screen:



View Profile Screen:



International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 17, Issue 2, 2024

8.CONCLUSION

One of the significant impediments with the progression of technology and medicine is the early detection of a disease, which is in this case, diabetes. However, in this study, systematic efforts were made into designing a model which is accurate enough in determining the onset of the disease. With the experiments conducted on the Pima Indians Diabetes Database, we have readily predicted this disease. Moreover, the results achieved proved the adequacy of the system, with an accuracy of 76% using the K-Nearest Neighbours classifiers. With this being said, it is hopeful that we can implement this model into a system to predict other deadly diseases as well. There can be room for further improvement for the automation of the analysis of diabetes or any other disease in the future.

In future, we will try to create a diabetes dataset in collaboration with a hospital or a medical institute and will try to achieve better results. We will be incorporating more Machine Learning and Deep learning models for achieving better results as well.

9.REFERENCES

[1] P. Saeedi, I. Petersohn, P. Salpea, B. Malanda, S. Karuranga,
N. Unwin, S. Colagiuri, L. Guariguata, A. A. Motala, K. Ogurtsova,
J. E. Shaw, D. Bright, and R.Williams, “Global and regional diabetes
prevalence estimates for 2019 and projections for 2030 and 2045:
Results from the international diabetes federation diabetes atlas,
9th edition,” Diabetes Research and Clinical Practice, vol. 157, p.
107843, 2019.

[2] A. Mir and S. N. Dhage, “Diabetes disease prediction using machine
learning on big data of healthcare,” in 2018 Fourth International
Conference on Computing Communication Control and Automation
(ICCUBEA), 2018, pp. 1–6.

[3] D. Sisodia and D. S. Sisodia, “Prediction of diabetes using
classification algorithms,” Procedia Computer Science, vol.

ISSN 2319-5991 www.ijerst.com
Vol. 17, Issue 2, 2024

132, pp. 1578 – 1585, 2018, international Conference on
Computational Intelligence and Data Science. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877050918308548

[4] J. Smith, J. Everhart, W. Dickson, W. Knowler, and R. Johannes,
"Using the adap learning algorithm to forcast the onset of diabetes
mellitus," Proceedings - Annual Symposium on Computer Applications
in Medical Care, vol. 10, 11 1988.

[5] P. S. Kohli and S. Arora, "Application of machine learning in disease
prediction," in 2018 4th International Conference on Computing
Communication and Automation (ICCCA), 2018, pp. 1–4.

[6] Wes McKinney, "Data Structures for Statistical Computing in
Python," in Proceedings of the 9th Python in Science Conference,
St´efan van der Walt and Jarrod Millman, Eds., 2010, pp. 56 – 61.

[7] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers,
P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J.
Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk,S
M. Brett, A. Haldane, J. F. del R'ıo, M. Wiebe, P. Peterson,
P. G'erard-Marchant, K. Sheppard, T. Reddy, W. Weckesser,
H. Abbasi, C. Gohlke, and T. E. Oliphant, "Array programming
with NumPy," Nature, vol. 585, no. 7825, pp. 357–362, Sep. 2020.
[Online]. Available: https://doi.org/10.1038/s41586-020-2649-2

[8] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion,
O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg,
J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and
Edouard Duchesnay, "Scikit-learn: Machine Learning in Python,"
Journal of Machine Learning Research, vol. 12, no. 85, p. 28252830,
2011.