Section A-Research paper
ISSN 2063-5346

Disease risk estimation of early stage Hyperglycemia using machine learning techniques

Abdul Rehman[1] Mr. Moksud Alam Mallik[2]

[1]Research Scholar, Dept. of Computer Science and Engineering, Lords Institute of Engineering & Technology, Hyderabad, Telangana

[2]Associate Professor, Dept. of Computer Science and Engineering, Lords Institute of Engineering & Technology, Hyderabad, Telangana

Abstract - Diabetes is an illness caused because of high glucose level in a human body. Diabetes should not be ignored if it is untreated then Diabetes may cause some major issues in a person like: heart related problems, kidney problem, blood pressure, eye damage and it can also affects other organs of human body. Diabetes can be controlled if it is predicted earlier. To achieve this goal this project work we will do early prediction of Diabetes in a human body or a patient for a higher accuracy through applying, Various Machine Learning Techniques. Machine learning techniques Provide better result for prediction by con- structing models from datasets collected from patients. 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), Decision Tree (DT), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest (RF), and naive bayes classifier have been used in this study. Among all these machine learning classifiers, Random Forest (RF) classifier has showed the best accuracy of almost 100%.

I. 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 type1 diabetes [1]. The most prevalent form of diabetes is type 2 diabetes, which is known as insulin-dependent or adultonset 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 [2] [3].

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 these 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.

II. LITERATURE REVIEW

K.VijiyaKumar et al. [1] proposed random Forest algorithm for the Prediction of diabetes develop a system which can perform early prediction of diabetes for a patient with a higher accuracy by using Random Forest algorithm in machine learning technique. The proposed model gives the best results for diabetic prediction and the result showed that the prediction system is capable of predicting the diabetes disease effectively, efficiently and most importantly, instantly. Nonso Nnamoko et al. [3] presented predicting diabetes onset: an ensemble supervised learning approach they used five widely used classifiers are employed for the ensembles and a meta-classifier is used to aggregate their outputs. The results are presented and compared with similar studies that used the

same dataset within the literature. It is shown that by using the proposed method, diabetes onset prediction can be done with higher accuracy. Tejas N. Joshi et al. [2] presented Diabetes Prediction Using Machine Learning Techniques aims to predict diabetes via three different supervised machine learning methods including: SVM, Logistic regression, ANN. This project proposes an effective technique for earlier detection of the diabetes disease. Deeraj Shetty et al. [5] proposed diabetes disease prediction using data mining assemble Intelligent Diabetes Disease Prediction System that gives analysis of diabetes malady utilizing diabetes patient's database. In this system, they propose the use of algorithms like Bayesian and KNN (K-Nearest Neighbor) to apply on diabetes patient's database and analyze them by taking various attributes of diabetes for prediction of diabetes disease. Muhammad Azeem Sarwar et al. [6] proposed study on prediction of diabetes using machine learning algorithms in healthcare they applied six different machine learning algorithms Performance and accuracy of the applied algorithms is discussed and compared. Comparison of the different machine learning techniques used in this study reveals which algorithm is best suited for prediction of diabetes. Diabetes Prediction is becoming the area of interest for researchers in order to train the program to identify the patient are diabetic or not by applying proper classifier on the dataset. Based on previous research work, it has been observed that the classification process is not much improved. Hence a system is required as Diabetes Prediction is important area in computers, to handle the issues identified based on previous research.

III. PROPOSED METHODOLOGY

Goal of the paper is to investigate for model to predict diabetes with better accuracy. We experimented with different classification and ensemble algorithms to predict diabetes. In the following, we briefly discuss the phase.

A. Dataset Description- the data is gathered from UCI repository which is named as Pima Indian Diabetes Dataset. The dataset have many attributes of 768 patients.

Distribution of Diabetic patient- We made a model to predict diabetes however the dataset was slightly imbalanced having around 500 classes labeled as 0 means negative means no diabetes and 268 labeled as 1 means positive means diabetic.

B. Data Preprocessing- Data preprocessing is most important process. Mostly healthcare related data contains missing vale and other impurities that can cause effectiveness of data. To improve quality and effectiveness obtained after mining process, Data preprocessing is done. To use Machine Learning Techniques on the dataset effectively this process is essential for accurate result and successful prediction. For Pima Indian diabetes dataset we need to perform pre processing in two steps.

1). Missing Values removal- Remove all the instances that have zero (0) as worth. Having zero as worth is not possible. Therefore this instance is eliminated. Through eliminating irrelevant features/instances we make feature subset and this process is called features subset selection, which reduces diamentonality of data and help to work faster.

2). Splitting of data- After cleaning the data, data is normalized in training and testing the model. When data is spitted then we train algorithm on the training data set and keep test data set aside. This training process will produce the training model based on logic and algorithms and values of the feature in training data. Basically aim of normalization is to bring all the attributes under same scale.

C. Apply Machine Learning- When data has been ready we apply Machine Learning Technique. We use different classification and ensemble techniques, to predict diabetes. The methods applied on Pima Indians diabetes dataset. Main objective to apply Machine Learning Techniques to analyze the performance of these methods and find accuracy of them, and also been able to figure out the responsible/important feature which play a major role in prediction. The Techniques are follows

1) Support Vector Machine- Support Vector Machine also known as svm is a supervised machine learning algorithm. Svm is most popular classification technique. Svm creates a hyperplane that separate two classes. It can create a hyperplane or set of hyperplane in high dimensional space. This hyper plane can be used for classification or regression also. Svm differentiates instances in specific classes and can also classify the entities which are not supported by data. Separation is done by through hyperplane performs the separation to the closest training point of any class. Algorithm- • Select the hyper plane which divides the class better. • To find the better hyper plane you have to calculate the distance between the planes and the data which is called Margin. • If the distance between the classes is low then the chance of miss conception is high and vice versa. So we need to • Select the class which has the high margin. Margin = distance to positive point + Distance to negative point.

2) K-Nearest Neighbor - KNN is also a supervised machine learning algorithm. KNN helps to solve both the classification and regression problems. KNN is lazy prediction technique.KNN assumes that similar things are near to each other. Many times data points which are similar are very near to each other.KNN helps to group new work based on similarity measure.KNN algorithm record all the records and classify them according to their similarity measure. For finding the distance between the points uses tree like structure. To make a prediction for a new data point, the algorithm finds the closest data points in the training data set — it's nearest neighbors. Here K= Number of nearby neighbors, it's always a positive integer. Neighbor's value is chosen from set of class. Closeness is mainly de

III. PROJECT EXECUTION

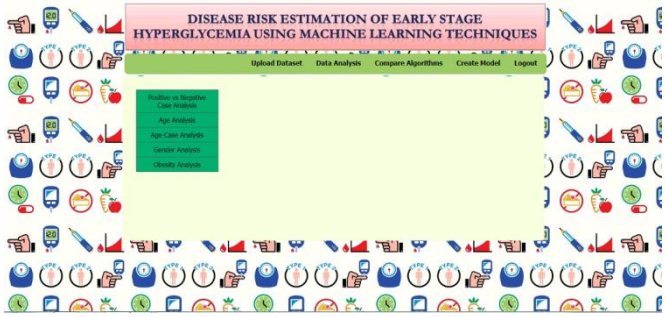
Home page:

Admin Login:

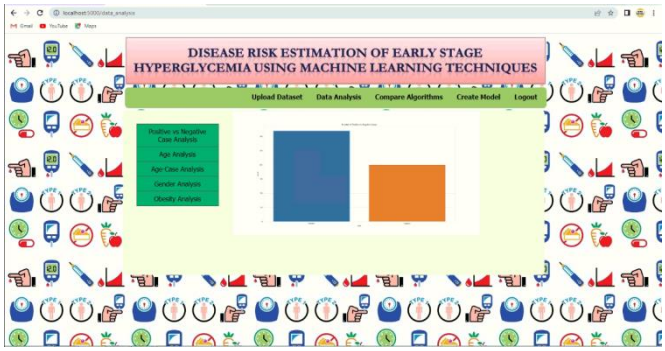
Upload Dataset:

Uploaded Dataset File Successfully

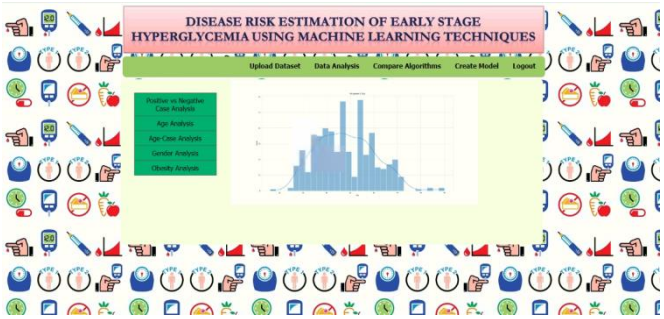
Data Analysis:



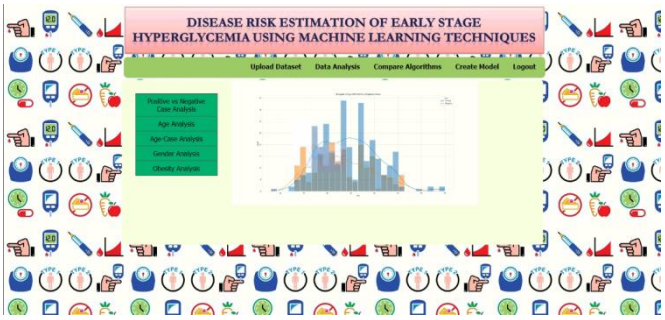
Positive vs Negative Case Analysis:



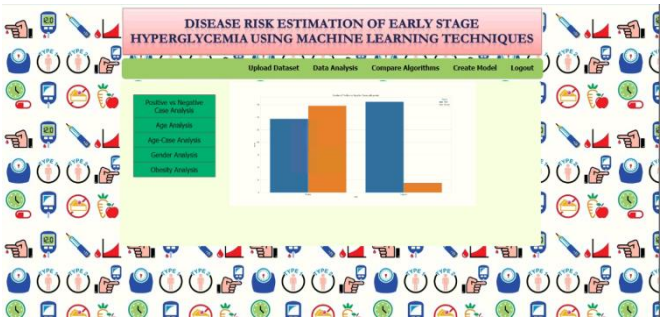
Age Analysis:



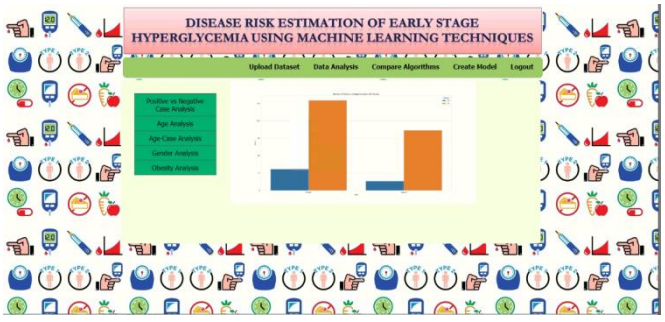
Case Analysis:



Gender Analysis:



Obesity Analysis:

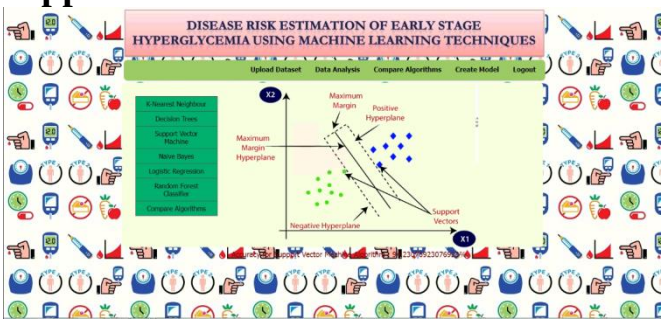


C

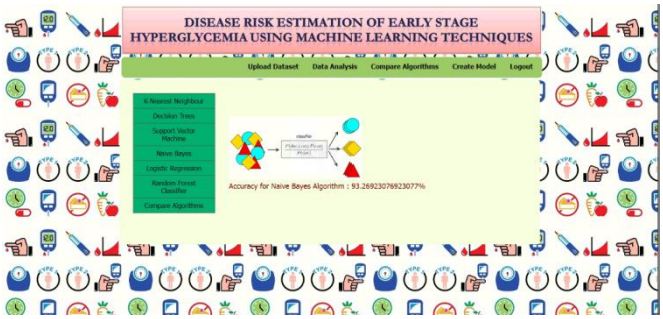
ompare Algorithms:

K-Nearest Neighbor:

Support Vector Machine:

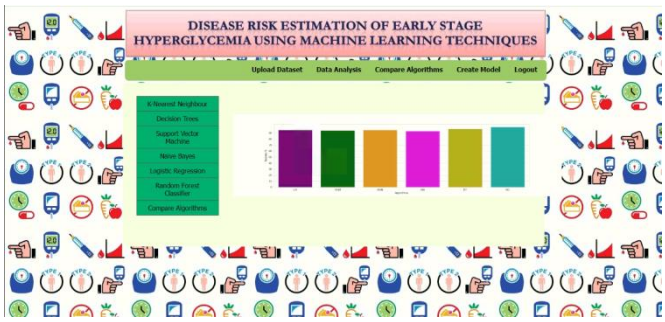


Naive Bayes:

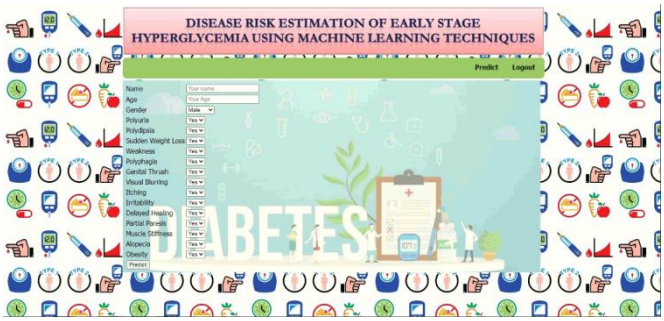


Decision Trees:

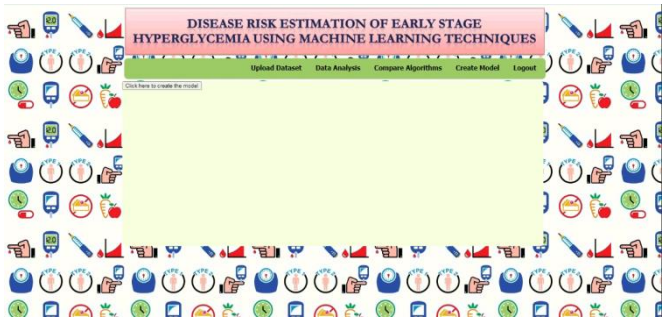
Compare Algorithms:

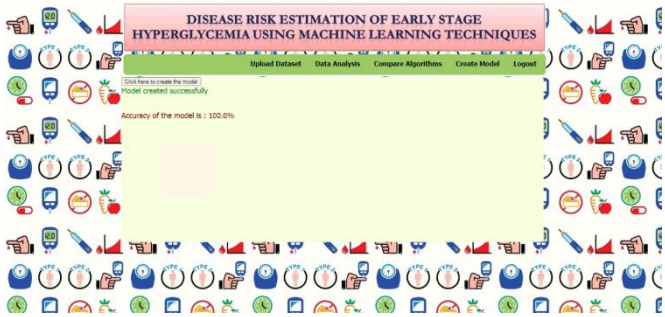


Predict:

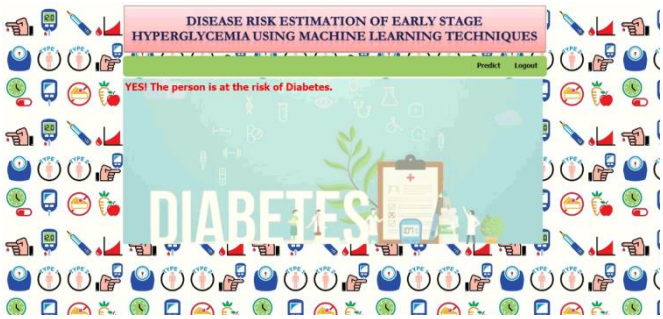


Create Model:





Result page:



User Login:

CONCLUSION

The main aim of this project was to design and implement Diabetes Prediction Using Machine Learning Methods and Performance Analysis of that methods and it has been achieved successfully. The proposed approach uses various classification and ensemble learning method in which Logistic Regression (LR), Decision Tree (DT), K-Nearest Neighbors

(KNN), Support Vector Machine (SVM), Random Forest (RF), and naive bayes classifier have been used in this study. Among all these machine learning classifiers, Random Forest (RF) classifiers are used. And almost 100% classification accuracy has been achieved. The Experimental results can be asst health care to take early prediction and make early decision to cure diabetes and save humans life. 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 patientsonly. In the latter study, we shall accumulate data from a variety of regions of the country. Aggregating different classifiers sometimes enhances the performances of prediction. We can apply the RF 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.

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.