Biomedical & Pharmacology Journal, June 2024. Vol. **17**(2), p. 1307-1316

Diabetes Prediction Using Machine Learning and Flask

N. Kushal Kumar Raju[1]*, Keshav Krishnamurthy[1], Bhuvanagiri Prahal Bhagavath[2], Nathan Shankar[2], A.M. Janani[3], N. Avinash[2], Aditya Ray[1] and P. Mahalakshmi[4]

[1]Electronics and Instrumentation Engineering, School of Electrical Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu, India.
[2]Electrical and Electronics Engineering, School of Electrical Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu, India.
[3]Computer Science and Engineering, Sri Sairam Engineering College, Chennai, Tamil Nadu, India.
[4]School of Electrical Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu, India.
*Corresponding Author E-mail: pmahalakshmi@vit.ac.in

https://dx.doi.org/10.13005/bpj/2944

(Received: 06 June 2022; accepted: 06 October 2023)

Diabetes is one of the costliest chronic diseases, it is a metabolic disorder in which a patient has excessive blood sugar levels due to the body's inability to create enough insulin, and it can also cause long-term harm to the heart, blood vessels, eyes, kidneys, and nerves. Adults with diabetes are twice as likely as non-diabetics to have a heart attack or stroke. Despite its massive impact on the global population, no kind of diabetes has a cure. Although most medications help patients manage their symptoms to some extent, diabetics nevertheless suffer several long-term health concerns. So, if we are able to predict diabetes early, we could control it and it can be done by using Machine learning techniques. Our work aim is to predict if the patient has diabetes using Machine learning techniques and the ensemble method. We will be using four algorithms which are SVM, KNN, Logistic Regression, and Random Forest classifier and we would also compare all four models to check which model is giving the best accuracy and link our best model to a web app that could predict if the patient has any chances of having diabetes.

Keywords: Ensemble; KNN; Logistic Regression; Machine learning; Random Forest; SVM.

Diabetes is a condition caused by a rise in blood glucose levels. Diabetes is a chronic condition that has caused a global healthcare catastrophe. Diabetes impacted around 463 million people worldwide in 2019, accounting for approximately 8.8 percent of the total adult population[1]. The healthcare sector is facing hurdles in crucial areas like electronic record management, computer-aided diagnosis, and disease predictions due to the need to lower healthcare costs and the shift to personalized healthcare. To meet these issues, machine learning provides a wide range of tools, methods, and systems[2]. According to the statistics of researchers, the occurrences of diabetes in men and women are similar[3]. Several experimental studies indicate that occurrences will continue to climb[1]. Diabetes is a significant chronic condition in children and teenagers as well. Young diabetics are at risk of dying because of the disease's acute consequences.[4] In 2019, diabetes

claimed the lives of almost 4.2 million people. It is the seventh leading cause of death in the globe[5,6]. Diabetes-related health expenditures were predicted to cost around $727 billion globally in the year 2017[1]. In the United States alone, the cost of diabetes in 2017 was close to $327 billion[7]. The average medical expense for a diabetic is 2.3 times more than for a non-diabetic[8]. Diabetes is a serious health concern because of its rising incidence, the rise of its complications as a leading cause of early illness and death, furthermore the tremendous and escalating cost it imposes on healthcare systems[9]. Serious consequences can include diabetic ketoacidosis, hyperosmolar hyperglycemia, and even death[10]. Severe long-term complications include cardiovascular disease, stroke, chronic renal disease, foot ulcers, nerve damage, vision loss, and cognitive impairment[4,11]. Big Data and the Cloud are two examples of new technologies that are helping to solve healthcare issues[12]. Predictive analytics strives to improve healthcare outcomes by accurately detecting diseases, improving patient care, and maximizing resources[13]. The complexity of updating the healthcare industry's tendency toward processing massive health data and accessing them for analysis and action will increase significantly. Because Big Data in the healthcare business is unstructured, it is necessary to structure and emphasize its magnitude into a nominal value using a realistic solution[14]. Diabetes has afflicted about 246 million individuals globally, with women accounting for the majority of those affected. According to WHO research, this figure is predicted to climb to more than 380 million by 2025. The illness has been ranked the fifth-deadliest disease in the United States, and there is little hope of treatment in the near future[15]. By 2035, it is anticipated that around 600 million individuals would have diabetes. Diabetes can be diagnosed using a variety of traditional approaches based on physical and chemical testing. However, early prediction of diabetes is a quite challenging task for medical practitioners due to complex interdependence on various factors as diabetes affects human organs such as kidneys, eyes, heart, nerves, foot, etc. Methods of data science have the potential to aid other scientific domains by throwing fresh light on prevalent topics. One such task is to assist in the prediction of medical data. To diagnose Diabetes, medical professionals require a trustworthy prediction methodology[16]. Employing various supervised learning methods of artificial neural networks, a study was conducted on the prediction of diabetes. Many diabetics between the ages of 25 to 78 provided data for the network's training. To verify accurate prediction, the optimal algorithm's prediction accuracy is computed[19]. Sugar from the foods we eat comes, and insulin is the hormone that helps the entry of sugar into the cells in order to give energy[21]. Machine learning is an emerging scientific field in data science dealing with how machines learn from experience. Filter algorithms are broad preprocessing methods that do not rely on a particular categorization method[25]. This project aims to develop a system that can perform an early prediction of diabetes for a patient with higher accuracy by combining the results of different machine learning techniques and integrating them with a web app so that the user can check their chances of having diabetes live.

MATERIALS AND METHODS

Dataset

The dataset generated was retrieved from Sylhet Diabetes Hospital patients in Sylhet, Bangladesh, and was approved by a doctor[22]. Utilizing large information examination one can study immense datasets and track down secret data[28]. The dataset contains 520 patients.

Attribute information

- Polyuria: Production of abnormally large volumes of dilute urine.
- Polydipsia: abnormally great thirst
- Sudden weight loss: reduction in your overall weight
- Weakness: feeling weak and tired
- Polyphagia: extreme hunger
- Genital thrush: yeast infection caused in private parts
- Visual blurring: lack of sharpness in your vision or having a blurred vision
- Itching: feeling itchiness in your body
- Irritability: being annoyed or irritated frequently
- Delayed healing: the wound has trouble healing or staying closed
- Partial paresis: feeling like mild paralysis, Unlike paralysis, people with paresis can still move their muscles
- Muscle stiffness: When your muscles feel tight

and you find it more difficult to move than you usually do, especially after rest

• Alopecia: having a partial or complete absence of hair from areas of the body where it normally grows

• Obesity: if your BMI is above 30 you are considered to be obese

• Class: the outcome of the patient, if he is positive or negative for diabetes

Data Preprocessing

The most crucial procedure is data pre-processing. Most healthcare-related data has missing values and other contaminants, which

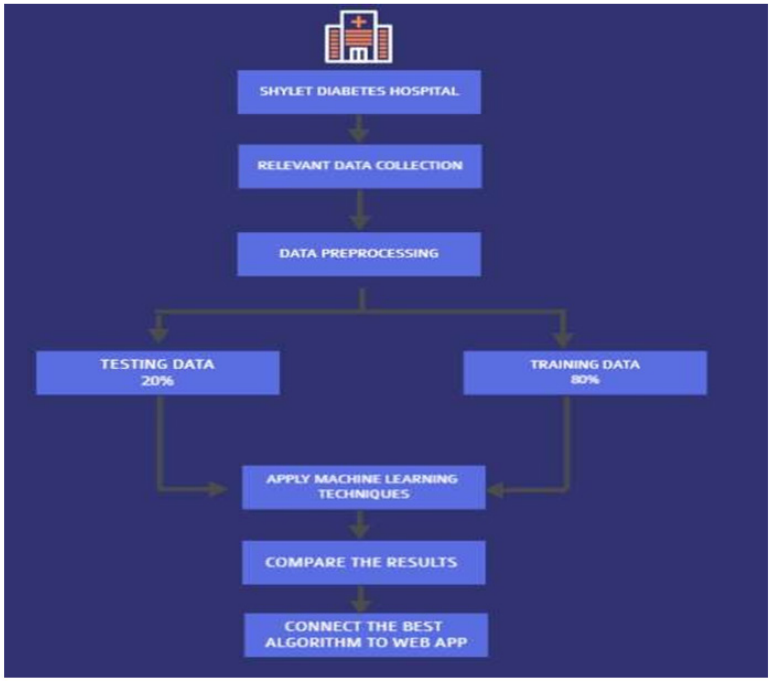

Fig. 1. Flowchart

Age	0
Gender	0
Polyuria	0
Polydipsia	0
sudden weight loss	0
weakness	0
Polyphagia	0
Genital thrush	0
visual blurring	0
Itching	0
Irritability	0
delayed healing	0
partial paresis	0
muscle stiffness	0
Alopecia	0
Obesity	0
class	0
dtype: int64	

Fig. 2. Checking for null values

can reduce data effectiveness. Data mining is capable of extracting hidden insights from massive amounts of diabetes-related data[27]. It employs a number of ways of analyzing massive volumes of data in order to find hidden knowledge[33]. It is the rigorous process of identifying instances in massive data sets that includes approaches at the intersection of artificial intelligence, machine learning, insights, and database system[31]. It is a collection of heuristics and computations used to extract a data mining model from data[18]. Medical data mining is used in knowledge collection and analysis to turn information gathered from research papers, medical reports, flow charts, and evidence tables into meaningful information for decision-making[17]. Data analytics is the act of evaluating and detecting hidden patterns in massive amounts of data in order to derive conclusions[20]. Big Data Analytics is important in the healthcare industry. Databases in the healthcare industry are huge in size. Using big data analytics, one may investigate

massive datasets to uncover hidden information and trends in order to gain knowledge from the data and forecast results accordingly[29]. Data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or data set, and it refers to distinguishing incomplete, incorrect, inaccurate, or tangential parts of the knowledge by substituting, modifying, or deleting the dirty or coarse data[32]. To improve the quality and effectiveness obtained after the mining process and data cleaning, Data preprocessing is done. To use Machine Learning Techniques on the dataset effectively this process is essential for accurate results and successful prediction.

Apply Machine Learning

Data can be categorized based on their characteristics. Classification is accomplished by creating a model based on existing records and sample data. One of the goals of categorization is to improve the consistency of the data-based outcomes[35]. When data is ready, Machine Learning Techniques are used. We use different classification and ensemble techniques, to predict diabetes. Manipulation of the input data provided to a single

	Age	Gender	Polyuria	Polydipsia	sudden weight loss	weakness	Polyphagia	Genital thrush	visual blurring	Itching	Irritability	delayed healing	partial paresis	muscle stiffness	Alopecia	Obesity
0	40	0	0	1	0	1	0	0	0	1	0	1	0	1	1	1
1	58	0	0	0	0	1	0	0	1	0	0	0	1	0	1	0
2	41	0	1	0	0	1	1	0	0	1	0	1	0	1	1	0
3	45	0	0	0	1	1	1	1	0	1	0	1	0	0	0	0
4	60	0	1	1	1	1	1	0	1	1	1	1	1	1	1	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
515	39	1	1	1	1	0	1	0	0	1	0	1	1	0	0	0
516	48	1	1	1	1	1	1	0	0	1	1	1	1	0	0	0
517	58	1	1	1	1	1	1	0	1	0	0	0	1	1	0	1
518	32	1	0	0	0	1	0	0	1	1	0	1	0	0	1	0
519	42	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Fig. 3. Replacing yes and no with 0s and 1s

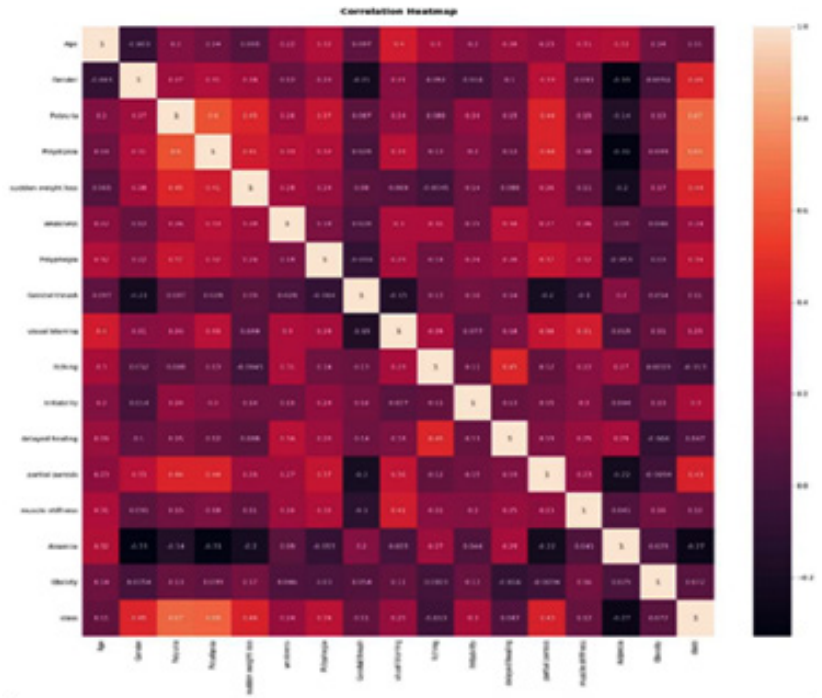

Fig. 4. Correlation Heat map

classifier is a typical method for constructing ensembles. This may be accomplished by running the classifier using a training set consisting of a randomly selected sample with a replacement from the original dataset[30]. The main objective to apply Machine Learning Techniques is to analyze the performance of these methods and find accuracy of them, and also be able to figure out the responsible/ important feature which plays a major role in prediction.

Web Application

Create a web app using Flask and connect it to our model which gave the best accuracy. We will save our best prediction model to a file using a library called Pickle. The model is then packaged into a web service that, when supplied data through a POST request, gives the diabetes prediction probability as a response. We will utilize the Flask web framework for this, which is a popular

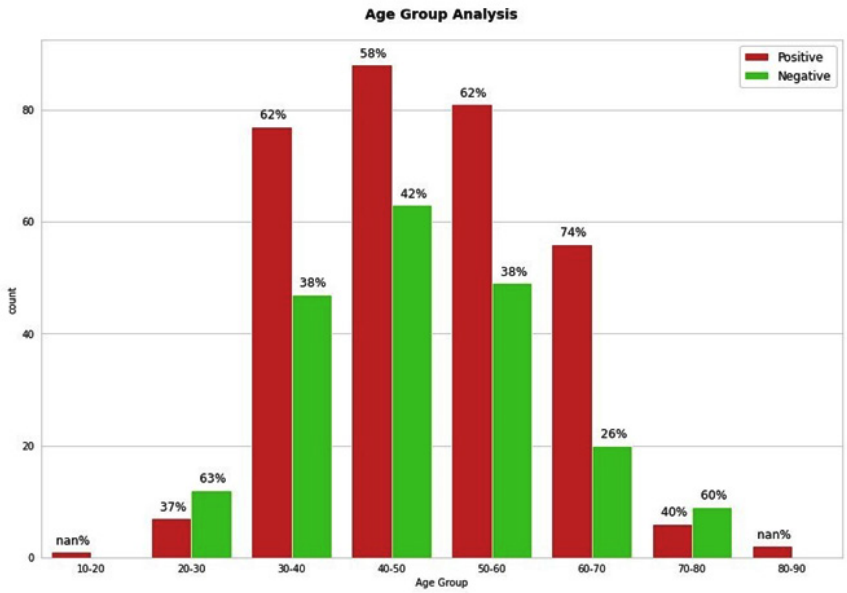

Fig. 5. Age group Data Analysis

Fig. 6. Web app

lightweight framework for constructing web services in Python. First, we'll utilize the request method to collect data from the user and save it in the appropriate variables. The model will then be used to predict the possibility that an individual is diabetic using 'predict()'. Then, based on the prediction, we will render the index.html page and display the necessary output.

Algorithms Used

• KNN- This technique be used for both classification and regression. This algorithm's principal function is to categorize a new object based on characteristics and training examples[34]. According to the k training samples, which are the closest nearby neighbors to the test person, this algorithm predicts the test sample's level and obtains the result with the greatest level possibility[23].

• LOGISTIC REGRESSION- Logistic Regression is a classification approach based on the idea of the probability that may be used as a predictive analytic methodology. It is used to solve categorization problems. This technique yields a discrete binary result between 0 and 1[24].

• SVM- SVM is a supervised machine learning technique that outperforms other machine learning algorithms and has been extensively tested in real-world classification problems and nonlinear function forecasting tasks[36].

• RANDOM FOREST- For bootstrapping and ensemble composition, random forest is a strategy that can avoid the overfitting problem[26].

Implementation

• In Figure 2. We have used isna().sum() which would sum all the null values present in our dataset, since the sum for all is 0 so there are no null values present in our dataset.

• We are replacing all "yes" and "no" to 1s and 0s, positive and negative outcomes with 1s and 0s and male and female with 0s and 1s respectively, to make it easy for our model to do data analysis

Data Preprocessing

Data Visualization and Analysis

• In Figure 4, We are using the heat map feature in the seaborns library to correlate our attributes. We can see that polyuria, polydipsia, sudden weight loss and gender show higher correlation with Class attribute which shows the outcome or target. So these attributes have higher weightage in terms of turning positive for diabetes.

• In Figure 5, Age group analysis is done to check which age group has more chances of diabetes, we can infer that age group 60-70 in the dataset has more chances of having diabetes.

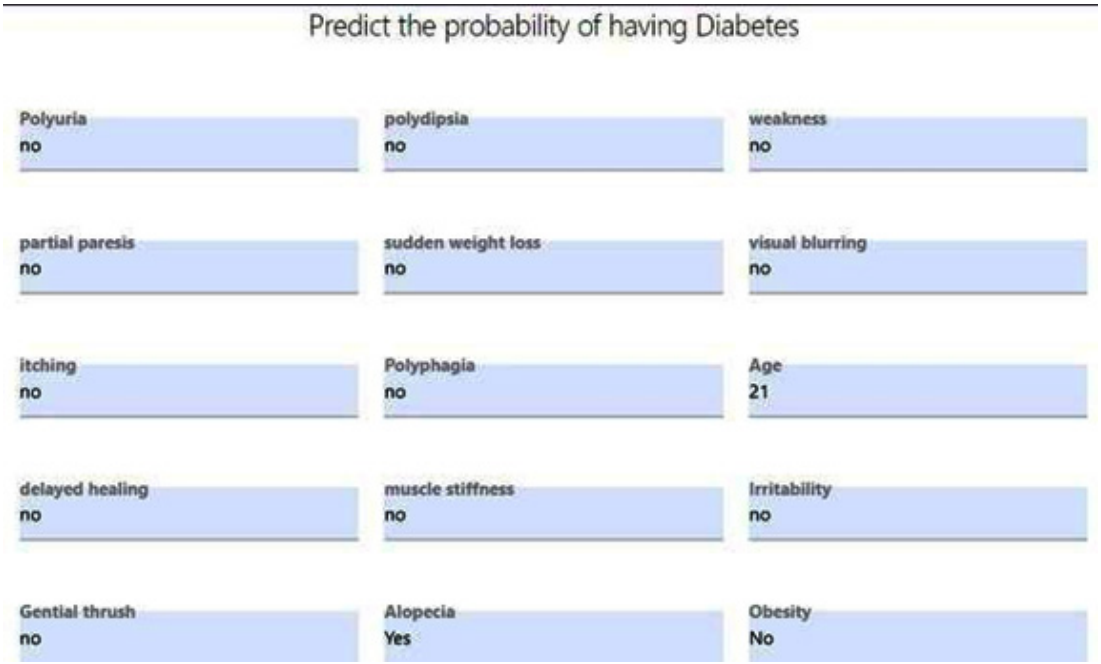

Fig. 7. Web app user inputs

Web App development

• Figure 6 depicts the Web application interface, and Figure 7 depicts the user inputs for all of the attributes that would be entered into our model to predict diabetes.

RESULTS AND DISCUSSION

In Figure 8, Classification report, model test score and confusion matrix are shown for each model. The Classification report and confusion matrix are used to evaluate the performance of all our models, Classification report shows the precision, recall, F1 Score, and support of our trained model. The ratio of genuine positives to the total of true and false positives is defined as precision. The ratio of true positives to the sum of true positives and false negatives is known as recall. The weighted harmonic mean of accuracy and recall is used to get the F1 score. The closer the F1 score number is to 1.0, the higher the model's projected performance. The number of actual instances of the class in the dataset is referred to as support. It does not differ between models; it just diagnoses the process of performance evaluation. So based on the results obtained, we can see that Random forest algorithm has shown the best performance since its f1-score, precision, and recall is 1.00.

Confusion Matrix

In Figure 9. The confusion matrix of all 4 models has been plotted using heatmap and seaborns library to compare the performance of each model. From the above results obtained we can see that False Negative for the random forest is 0.00% which means when a person is not diabetic, it did not show he is diabetic.

• Table 1 shows the comparison of all our models and the Random Forest model showed the highest accuracy.

Cross validation

• In Figure 10, Cross-validation is done for all 4 models to validate our models, to estimate how

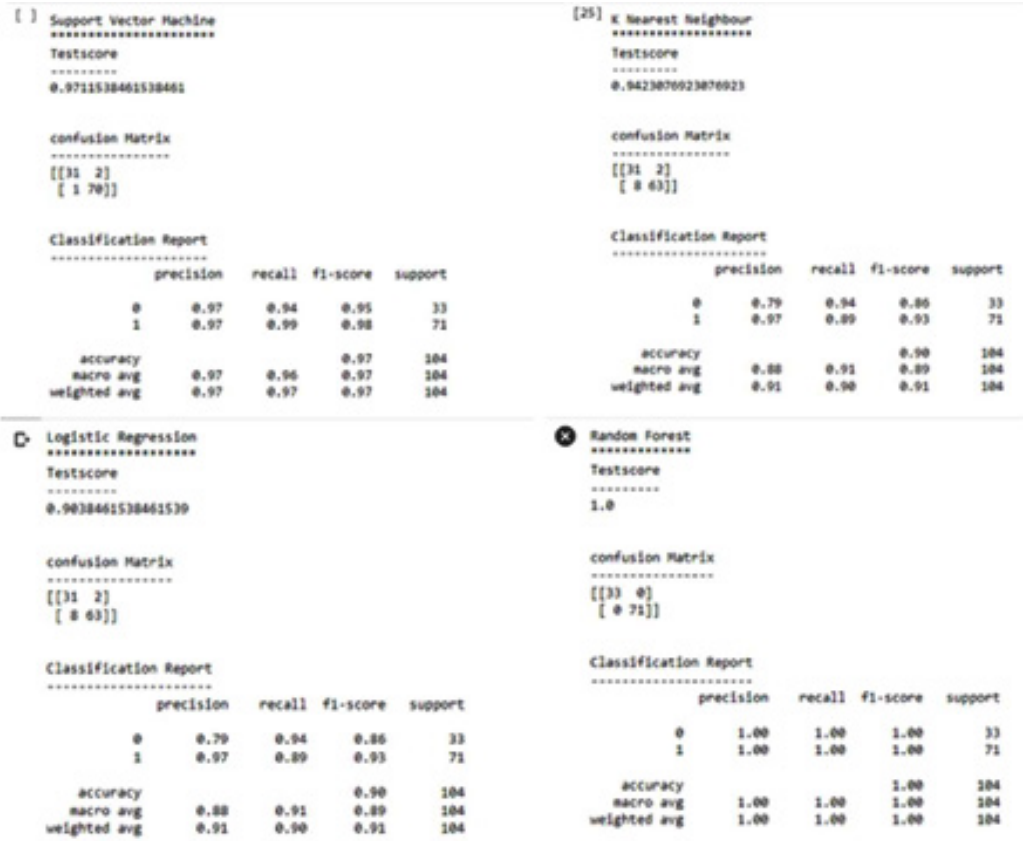

Fig. 8. Classification report of all models

accurately a predictive model will perform in practice

Web App

• In Figure 11, the result of the patient's diabetes is shown in the Web application. The cutoff value for the prediction of diabetes is taken as 0.5, so if a patient gets a probability of more than 0.5 then the patient is at risk of having diabetes.

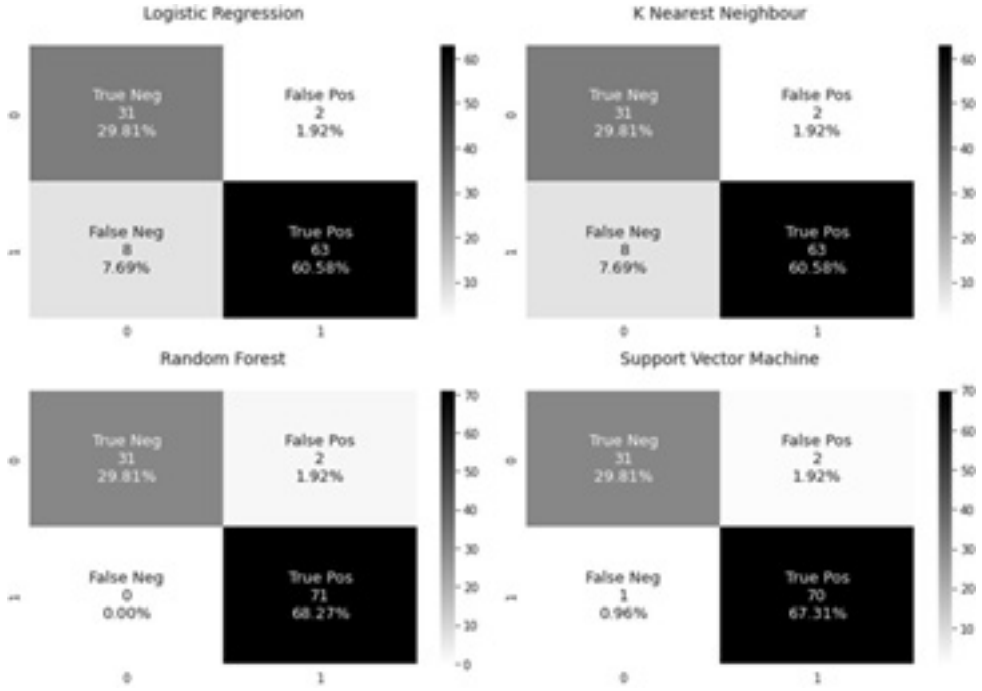

Fig. 9. Confusion matrix

```
Logistic Regression
******************************
Min:0.7884615384615384
Max:1.0
Avg:0.9326923076923077

K Nearest Neighbour
******************************
Min:0.8846153846153846
Max:1.0
Avg:0.9692307692307691

Random Forest
******************************
Min:0.9038461538461539
Max:1.0
Avg:0.9807692307692308

Support Vector Machine
******************************
Min:0.8653846153846154
Max:1.0
Avg:0.9692307692307693
```

Fig. 10. Cross-validation

CONCLUSION

• From the above table it can be observed that random forest gives the highest accuracy when trained with the dataset we have and we were also able to connect our model with a web app.

• If a person has diabetes but is predicted not to have it (false negative), this is the worst-case situation and might have devastating consequences. This is prevented in our case because the false negative is 0.0% for Random Forest.

Table 1. Comparision of all algorithm results

Algorithm	Accuracy (%)
Logistic Regression	93.27
Support Vector Machine	96.92
K-nearest neighbors	96.92
Random Forest	98.08

Fig.11. Web app result

• So we were able to develop a system that can perform early prediction of diabetes for a patient with higher accuracy by using machine learning technique.

ACKNOWLEDGEMENT

None.

Conflict of Interest

There is no conflict of interest.

Funding Sources

There is no funding Source.

REFERENCES

1. R. Thomas, S. Halim, S. Gurudas, S. Sivaprasad, and D. Owens, "Idf diabetes atlas: A review of studies utilising retinal photography on the global prevalence of diabetes related retinopathy between 2015 and 2018," Diabetes research and clinical practice, vol. 157, p. 107840,2019.
2. B. Nithya and Dr. V. Ilango," Predictive Analytics in Health Care Using Machine Learning Tools and Techniques", International Conference on Intelligent Computing and Control Systems, 978-1-5386-2745-7,2017.
3. T. Vos, A. D. Flaxman, M. Naghavi, R. Lozano, C. Michaud, M. Ezzati,K. Shibuya, J. A. Salomon, S. Abdalla, V. Aboyans et al., "Years lived with disability (ylds) for 1160 sequelae of 289 diseases and injuries1990–2010: a systematic analysis for the global burden of disease study 2010," The lancet, vol. 380, no. 9859, pp. 2163–2196, 2012.
4. W. H. Organization et al., "Diabetes fact sheet n 312. october 2013,"Archived from the original on, vol. 26, 2013.
5. U. Diabetes and H. Lobby, "What is diabetes," Diabetes UK, 2014.
6. W. H. Organization et al.,"The top 10 causes of death fact sheet no 310," Geneva, Switzerland: World Health Organization, 2013.
7. A. D. Association et al., "Economic costs of diabetes in the us in 2017,"Diabetes care, vol. 41, no. 5, pp. 917–928, 2018.
8. C. for Disease Control, Prevention et al., "National diabetes statistics report, 2020," Atlanta, GA: Centers for Disease Control and Prevention, US Department of Health and Human Services, pp. 12–15, 2020.
9. W.H.Organization etal., Guidelines for the prevention, management and care of diabetes mellitus, 2006.
10. A. E. Kitabchi, G. E. Umpierrez, J. M. Miles, and J. N. Fisher,"Hyperglycemic crises in adult patients with diabetes," Diabetes care,vol. 32, no. 7, pp. 1335–1343, 2009.
11. E. Saedi, M. R. Gheini, F. Faiz, and M. A. Arami, "Diabetes mellitus and cognitive impairments," World journal of diabetes, vol. 7, no. 17,p. 412, 2016.
12. P. Suresh Kumar and S. Pranavi "Performance Analysis of Machine Learning Algorithms on Diabetes Dataset using Big Data Analytics", International Conference on Infocom Technologies and Unmanned Systems, 978-1-5386-0514-1, Dec. 18-20, 2017.
13. Gauri D. Kalyankar, Shivananda R. Poojara and Nagaraj V. Dharwadkar," Predictive Analysis of Diabetic Patient Data Using Machine Learning and Hadoop", International Conference On I-SMAC,978-1-5090-3243-3,2017
14. Dr Saravana kumar N M, Eswari T, Sampath P and Lavanya S," Predictive Methodology for Diabetic Data Analysis in Big Data", 2nd International Symposium on Big Data and Cloud Computing,2015
15. Aiswarya Iyer, S. Jeyalatha and Ronak Sumbaly," Diagnosis of Diabetes Using Classification Mining Techniques", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.1, January 2015.
16. K. Rajesh and V. Sangeetha, "Application of Data Mining Methods and Techniques for Diabetes Diagnosis", International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 3, September 2012.
17. V. Kumar and L. Velide, "A data mining approach for prediction and treatment of diabetes disease," Int J Sci Invent Today, vol. 3, pp. 73–9, 2014.
18. P. Agrawal and A. Dewangan, "A brief survey on the techniques used for the diagnosis of diabetes-mellitus," Int. Res. J. of Eng. and Tech. IRJET, vol. 2, pp. 1039–1043, 2015.

19. M. A. Sapon, K. Ismail, and S. Zainudin, "Prediction of diabetes by Using artificial neural network," in Proceedings of the 2011 International Conference on Circuits, System and Simulation, Singapore, vol. 2829, 2011, p. 299303.
20. D. Singh, E. J. Leavline, and B. S. Baig, "Diabetes prediction using medical data," Journal of Computational Intelligence in Bioinformatics, vol. 10, no. 1, pp. 1–8, 2017.
21. T. M. Ahmed, "Developing a predicted model for diabetes type 2 treatment plans by using data mining," Journal of Theoretical and Applied Information Technology, vol. 90, no. 2, p. 181, 2016.
22. M. F. Islam, R. Ferdousi, S. Rahman, and H. Y. Bushra, "Likelihood prediction of diabetes at early stage using data mining techniques," in Computer Vision and Machine Intelligence in Medical Image Analysis. Springer, 2020, pp. 113–125.
23. S. Kaghyan and H. Sarukhanyan, "Activity recognition using k-nearest neighbor algorithm on smartphone with tri-axial accelerometer" in Inter-national Journal of Informatics Models and Analysis (IJIMA) ITHEA International Scientific Society, Bulgaria, vol. 1, pp. 146-156, 2012.
24. D. G. Kleinbaum, K. Dietz, M. Gail, M. Klein and M. Klein, Logistic regression, Springer, 2002.
25. T. M. Phuong, Z. Lin, and R. B. Altman, "Choosing snps using feature selection," in 2005 IEEE Computational Systems Bioinformatics Conference (CSB'05).
26. J. Ali, R. Khan, N. Ahmad, and I. Maqsood, "Random forests and decision trees," International Journal of Computer Science Issues (IJCSI),vol. 9, no. 5, p. 272, 2012.
27. Tejas N. Joshi, Prof. Pramila M. Chawan, "Diabetes Prediction Using Machine Learning Techniques".Int. Journal of Engineering Research and Application, Voljb. 8, Issue 1, (Part-II) January 2018, pp.-09-13
28. Ayush Anand and Divya Shakti," Prediction of Diabetes Based on Personal Lifestyle Indicators", 1st International Conference on Next Generation Computing Technologies, 978-1-4673-6809-4, September 2015.
29. Aishwarya Mujumdar, V Vaidehi, "Diabetes Prediction using Machine Learning Algorithms", Procedia Computer Science, Volume 165, 2019.
30. Nonso Nnamoko, Abir Hussain, David England, "Predicting Diabetes Onset: an Ensemble Supervised Learning Approach ". IEEE Congress on Evolutionary Computation (CEC), 2018.
31. Deeraj Shetty, Kishor Rit, Sohail Shaikh, Nikita Patil, "Diabetes Disease Prediction Using Data Mining ".International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), 2017.
32. K.VijiyaKumar, B.Lavanya, I.Nirmala, S.Sofia Caroline, "Random Forest Algorithm for the Prediction of Diabetes ".Proceeding of International Conference on Systems Computation Automation and Networking, 2019.
33. Mani Butwall and Shraddha Kumar," A Data Mining Approach for the Diagnosis of Diabetes Mellitus using Random Forest Classifier", International Journal of Computer Applications, Volume 120 - Number 8,2015
34. Utsha Das, Azmain Yakin Srizon† , Md. Ansarul Islam, Dhiman Sikder Tonmoy§ ,Md. Al Mehedi Hasan"Prognostic Biomarkers Identification for Diabetes Prediction by Utilizing Machine Learning Classifiers", 2020 2nd International Conference on Sustainable Technologies for Industry 4.0 (STI), 19-20 December, Dhaka.
35. Humar Kahramanli and Novruz Allahverdi, "Design of a Hybrid System for the Diabetes and Heart Disease", Expert Systems with Applications: An International Journal, Volume 35 Issue 1-2, July, 2008.
36. X. Yang, G. Zhang, J. Lu and J. Ma, "A kernel fuzzy c-means clustering-based fuzzy supportvector machine algorithm for classification problems with outliers or noises", IEEE Transactions on Fuzzy Systems, vol. 19, no. 1, pp. 105-115, 2010.