DIABETES PREDICTION USING MACHINE LEARNING TECHNIQUES ON AWS CLOUD

[1]N.Priyanka, [2]Pundra Raga Sree Reddy, [3]Kothapally Sai Sumanth,[4] Enukonda Harshavardhan Reddy

Assistant Professor in department Of IT Teegala Krishna Reddy Engineering College

priyanka.nelloju17@gmail.com

UG Scholars In Department of IT Teegala Krishna Reddy Engineering College

[2]ragasreereddy0@gmail.com ,[3] kothapallysaisumanth2@gmail.com ,[4] *enukondaharsha@gmail.com*

Abstract

ML has contributed towards opening up new frontiers in several sectors including health and medical facilities. Multiple ML approaches has been designed and developed to execute prediction oriented analysis on big data acquired from several devices. Performing analysis through prediction is critical and daunting yet eventually it can aid and assist medical professionals and healthcare providers to arrive at strategic and wise assessments that could prove to be effective during prognosis as well as diagnosis at the time of patients' treatments. This study takes into account the notion of analysis based on prediction especially in healthcare domain where ML algorithms are utilized to perform the necessary research. To validate and evaluate our proposed research, a training dataset comprising of patient's health records are acquired and 6 diverse ML algorithms are employed on it.Analysis of performance and effectiveness of the ML algorithms and their comparative efficiency in prediction of diabetes has been elaborated and discussed in detail. Evaluation of the various ML approaches helps towards understanding the suitability of reliable ML algorithm that could be utilized for predicting diabetes and the associated symptoms. This study is thus aimed at assisting medical practitioners and health personnel to detect the diabetes earlier through application of appropriate ML methods. These algorithms utilised are artificial neural networks(ANN) , XG boosting ,Ada boosting, K Nearest Neighbours (KNN), Support Vector Machine (SVM), Decision Tree (DT) . Through comparing and validating the above mentioned machine learning techniques facilitates the prediction of diabetes by means of an application where the users can enter the relevant details and acquire prediction based results.

I INTRODUCTION

Medical sector mainly deals with enormous amount of information that comprises of critically sensitive and private data which needs to be securely managed and safeguarded from un authorised access and modification. Owing to our sedentary and non healthy lifestyle Diabetes Mellitus is fast emerging and can be considered to be the most prevalent and one of the deadly diseases across the globe. Medical practitioners are in need of a robust and secure prediction scheme to accurately detect and identify Diabetes that could enable them to offer further customized health care facilities. Multiple machine learning oriented techniques are valuable in investigating the information from several perceptions and analyzing them to gain valuable insights that would validate the information to understand its relevance. Understanding the significance and relevance of specific information from easily available and highly accessible large volumes of data can offer necessary knowledge through appropriate application of data mining methodologies. The foremost objective is to verify and recognize patterns that exist among datasets and deduce a meaningful relationship from the acquired patterns to convey noteworthy and valuable information for the clients. Uncontrolled and undetected diabetes can lead to heart ail- ment, kidney infections, nerve problems, and can even lead to losing sight. Mining the diabetes information in a resourceful manner could prove to be a complex and decisive task as further analysis solely depends upon the initial information obtained.

The relevant approaches for performing data mining and processes involved needs to be determined to locate the suitable techniques to perform categorization of Diabetes dataset and derive noteworthy associations among them. In this research, ML based analysis is performed to achieve the accurate prediction of diabetes. The mining tool utilised is WEKA which is responsible for effective diagnosis of diabetes.

The Pima Indian diabetes records were obtained from UCI repository to conduct necessary analysis. The dataset was scrutinised and investigated to construct an efficient design framework that can aid the healthcare professionals to predict and diagnose diabetes disease. Our study involves application of bootstrapping resembling procedure to improve the predic tion accuracy and performance of our suggested method and then through administering ML algorithms like NB, DT and KNN to validate and analyze their effectiveness.

II LITERATURE SURVEY

Medical care using with Big Data Analytics

The introduction of Big Data Analytics (BDA) in healthcare will allow to use new technologies both in treatment of patients and health management. The paper aims at analyzing the possibilities of using Big Data Analytics in healthcare. The research is based on a critical analysis of the literature, as well as the presentation of selected results of direct research on the use of Big Data Analytics in medical facilities. The direct research was carried out based on research questionnaire and conducted on a sample of 217 medical facilities in Poland. Literature studies have shown that the use of Big Data Analytics can bring many benefits to medical facilities, while direct research has shown that medical facilities in Poland are moving towards data-based healthcare because they use structured and unstructured data, reach for analytics in the administrative, business and clinical area. The research positively confirmed that medical facilities are working on both structural data and unstructured data. The following kinds and sources of data can be distinguished: from databases, transaction data, unstructured content of emails and documents, data from devices and sensors. However, the use of data from social media is lower as in their activity they reach for analytics, not only in the administrative and business but also in the clinical area. It clearly shows that the decisions made in medical facilities are highly data-driven. The results of the study confirm what has been analyzed in the literature that medical facilities are moving towards data-based healthcare, together with its benefits.

Inthispaperweintroduceareal-timeobstacledetectionandclassificationsystemdesignedto assist visually impaired people to navigate safely, in indoor and outdoor environments,by handling a smartphone device. We start by selecting a set of interest points extractedfrom an image grid and tracked using the multiscale Lucas - Kanade algorithm. Then, weestimate the camera and background motion through a set of homographic transforms.Other types of movements are identified using an agglomerative clustering technique.Obstacles are marked as urgent or normal based on their distance to the subject and theassociated motion vector orientation. Following, the detected obstacles are fed/sent to anobject classifier. We incorporate HOG descriptor into the Bag of Visual Words (BoVW)retrieval framework and demonstrate how this combination may be used for obstacleclassification in video streams. The experimental results demonstrate that our approach is effective in image sequences with significant camera motion and achieves high accuracy rates, while being computational Efficient.

Enhancing the Data Security in Cloud

This paper presents Hybrid (RSA & AES) encryption algorithm to safeguard data security in Cloud. Security being the most important factor in cloud computing has to be dealt with great precautions. This paper mainly focuses on the following key tasks: 1. Secure Upload of data on cloud such that even the administrator is unaware of the contents. 2. Secure Download of data in such a way that the integrity of data is maintained. 3. Proper usage and sharing of the public, private and secret keys involved for encryption and decryption. The use of a single key for both encryption and decryption is very prone to malicious attacks. But in hybrid algorithm, this problem is solved by the use of three separate keys each for encryption as well as decryption. Out of the three keys one is the public key, which is made available to all, the second one is the private key which lies only with the user. In this way, both the secure upload as well as secure download of the data is facilitated using the two respective keys. Also, the key generation technique used in this paper is unique in its own way. This has helped in avoiding any chances of repeated.

III EXISTING SYSTEM

Data Quality and Representativeness:

Limited or biased data can affect the model's performance. If the training data doesn't accurately represent the diverse population of patients, the model may not generalize well.

Ethical and Privacy Concerns:

Handling sensitive health data requires strict adherence to privacy regulations. Ensuring that the system complies with ethical guidelines and legal requirements is crucial.

Interpretability of Models:

Some machine learning models, especially complex ones like neural networks, can be difficult to interpret. Understanding how and why a model makes a particular prediction is important, especially in healthcare where decisions can have significant consequences.

Generalization to New Cases:

The model's ability to generalize to new, unseen cases is crucial. Overfitting to the training data or lack of diversity in the dataset can hinder the model's performance on real-world cases.

Integration with Clinical Workflow:

For practical utility, the system needs to seamlessly integrate into the existing clinical workflow. If the integration is not smooth, healthcare professionals may be reluctant to adopt the technology.

Limited Explain ability:

Many advanced machine learning models, such as deep learning, lack inherent explain ability. Providing explanations for model predictions, especially in critical healthcare decisions, is important for gaining trust among medical professionals.

Scalability and Resource Utilization:

Scalability issues can arise, particularly when deploying the system in a real-world, cloud-based environment. Ensuring that the system can handle a growing volume of data and users is essential.

Model Maintenance and Updates:

Machine learning models require continuous monitoring, maintenance, and periodic updates. Failure to update the model with new data or adapt to changing healthcare conditions can lead to a decline in prediction accuracy over time.

Cost of Implementation:

Deploying and maintaining machine learning models on cloud platforms like AWS may involve costs. Understanding and managing these costs is essential for the sustainability of the system.

IV PROPOSED SYSTEM

The proposed system for the project titled "Machine Learning-based Diabetes Prediction using AWS Cloud" aims to address the limitations of existing systems and enhance the efficiency and effectiveness of diabetes prediction in a healthcare context. Building on the foundations of the current research, the proposed system envisions an improved data collection and preprocessing pipeline to ensure the quality and representativeness of the dataset. Advanced feature selection techniques will be employed to identify the most relevant factors for diabetes prediction, mitigating potential biases and improving the generalization of the models.

In the proposed system, a diverse set of machine learning algorithms, including Artificial Neural Networks (ANN), XG Boosting, Ada Boosting, K Nearest Neighbours (KNN), Support Vector Machine (SVM), and Decision Tree (DT), will be further refined and optimized. The emphasis will be on enhancing the interpretability of these models, addressing the challenge of understanding and explaining complex predictions in the healthcare domain. The system will also prioritize ethical considerations and privacy concerns, ensuring compliance with regulations governing the handling of sensitive health data.

To facilitate seamless integration into the clinical workflow, the proposed system will undergo rigorous testing and validation, assessing its performance on diverse patient populations and real-world scenarios. The goal is to create a scalable solution that leverages the capabilities of AWS cloud services, optimizing resource utilization and accommodating future growth in data volume and user interactions. Additionally, the system will incorporate user-friendly interfaces, allowing medical professionals to input relevant patient details and obtain reliable diabetes predictions in a user-friendly manner.

Continuous monitoring and maintenance protocols will be implemented to keep the machine learning models up-to-date with evolving healthcare conditions. Regular updates to the system will be deployed to adapt to changing data patterns, ensuring the sustainability and accuracy of diabetes predictions over time. Overall, the proposed system aspires to be a comprehensive and reliable tool for early diabetes detection, supporting medical practitioners and healthcare providers in making informed and timely decisions for patient care.

Advantages

Enhanced Prediction Accuracy:

The utilization of diverse machine learning algorithms, including Artificial Neural Networks (ANN), XG Boosting, Ada Boosting, K Nearest

Neighbours (KNN), Support Vector Machine (SVM), and Decision Tree (DT), contributes to a robust and accurate prediction model. By leveraging multiple algorithms, the system can capture complex patterns in the data, improving overall prediction accuracy.

Cloud-Based Scalability:

The integration of the system with AWS cloud services provides scalability, allowing the platform to handle an increasing volume of data and user interactions. This ensures that the system remains effective and responsive even as the dataset and user base grow over time.

User-Friendly Interface:

The proposed system includes a user-friendly interface that enables medical professionals to input relevant patient details easily. This user interface streamlines the process of obtaining predictions, making it accessible to healthcare providers who may not have extensive experience with machine learning technologies.

Ethical Compliance and Data Privacy:

The system prioritizes ethical considerations and data privacy, ensuring compliance with regulations governing the handling of sensitive health data. By implementing robust security

V IMPLEMENTATION

Data Collection and Preprocessing Module:

This module focuses on acquiring a diverse and representative dataset of patient health records. It includes processes for cleaning, handling missing values, and preprocessing the data to ensure its quality and suitability for machine learning model training.

Feature Selection and Optimization Module:

In this module, advanced feature selection techniques are implemented to identify the most relevant factors for diabetes prediction. The goal is to enhance the efficiency of the machine learning algorithms by focusing on key features and mitigating potential biases in the dataset. Optimization techniques are also applied to fine-tune the model parameters.

Machine Learning Algorithm Implementation Module:

This module incorporates the implementation of various machine learning algorithms, including Artificial Neural Networks (ANN), XG Boosting, Ada Boosting, K Nearest Neighbours (KNN), Support Vector Machine (SVM), and Decision Tree (DT). Each algorithm is trained on the preprocessed data to create prediction models for diabetes.

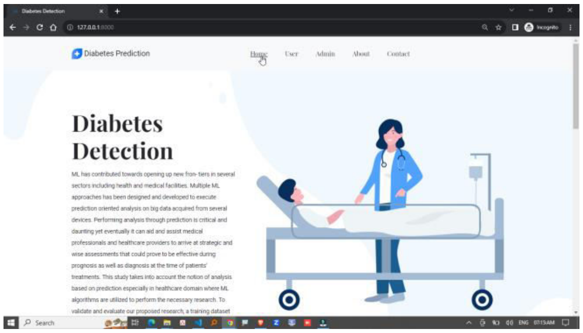
AWS Integration and Cloud Deployment Module:

The integration with AWS cloud services is a crucial aspect of the system. This module involves deploying the machine learning models on the AWS platform, leveraging services such as Amazon SageMaker for model training and hosting. This ensures scalability, resource optimization, and efficient utilization of cloud infrastructure.

User Interface and Interaction Module:

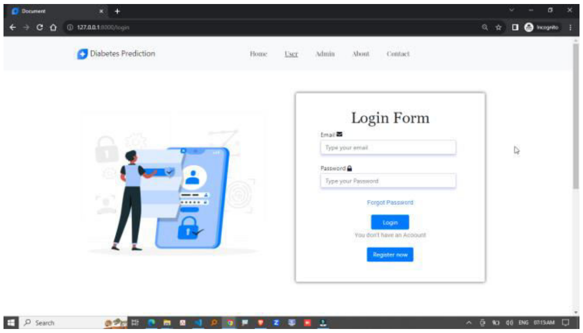
The user interface module is designed to create a user-friendly platform for medical professionals. It allows users to input relevant patient details and obtain diabetes predictions. This module ensures seamless interaction between the machine learning models and end-users, promoting easy adoption and integration into the clinical workflow.

VI RESULTS



Diabetes Detection

ML has contributed towards opening up new fron- tiers in several sectors including health and medical facilities. Multiple ML approaches has been designed and developed to execute prediction oriented analysis on big data acquired from several devices. Performing analysis through prediction is critical and daunting yet eventually it can aid and assist medical professionals and healthcare providers to arrive at strategic and wise assessments that could prove to be effective during prognosis as well as diagnosis at the time of patients' treatments. This study takes into account the notion of analysis based on prediction especially in healthcare domain where ML algorithms are utilized to perform the necessary research. To validate and evaluate our proposed research, a training dataset



Login Form

Email

Type your email

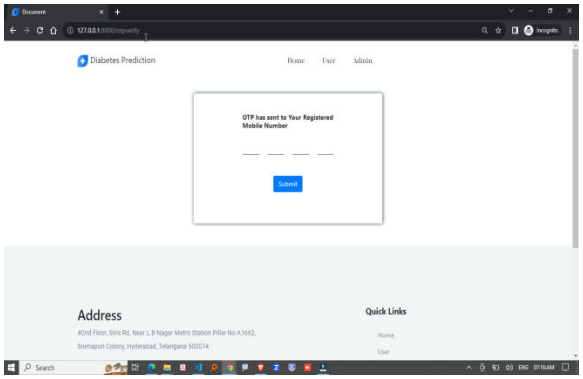
Password

Type your Password

Forgot Password

You don't have an Account

User's Login page

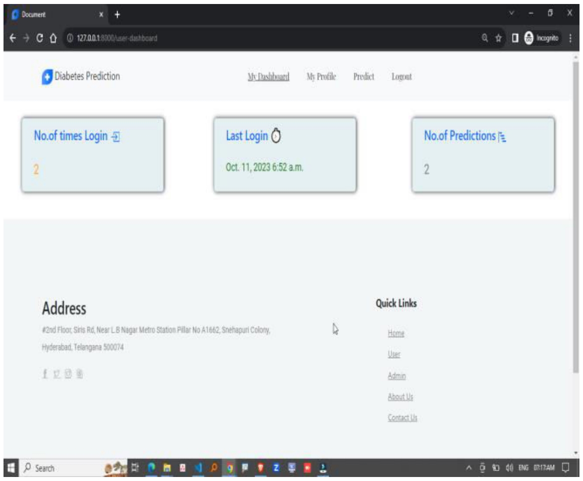


OTP has sent to Your Registered Mobile Number

Address

#2nd Floor, Siris Rd, Near L.B Nagar Metro Station Pillar No A1662, Snehapuri Colony, Hyderabad, Telangana 500074

User Mobile number Verification



No.of times Login 2

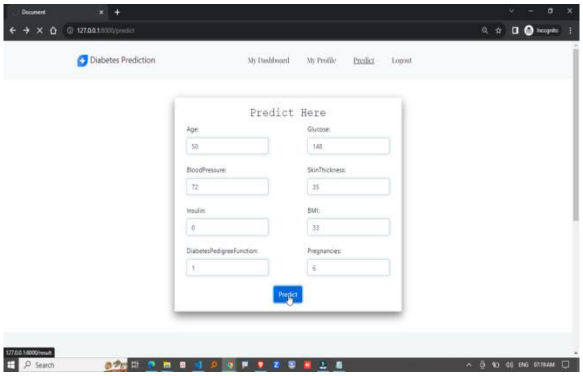
Last Login Oct. 11, 2023 6:52 a.m.

No.of Predictions 2

Address

#2nd Floor, Siris Rd, Near L.B Nagar Metro Station Pillar No A1662, Snehapuri Colony, Hyderabad, Telangana 500074

User's Dashboard



Predict Here

Age: 50

Glucose: 148

BloodPressure: 72

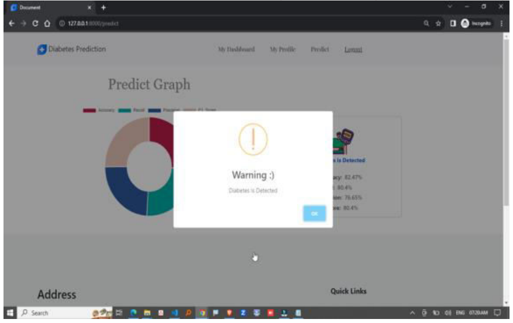
SkinThickness: 35

Insulin: 0

BMI: 33

DiabetesPedigreeFunction: 1

Pregnancies: 6

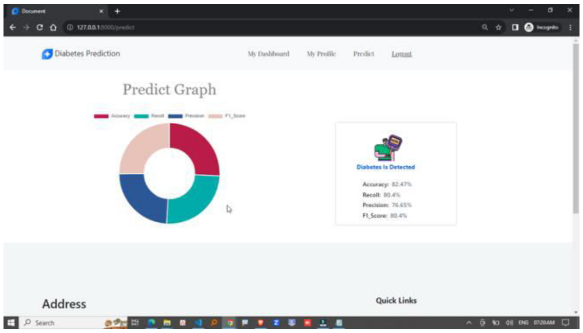


Warning :)

Diabetes is Detected

OK

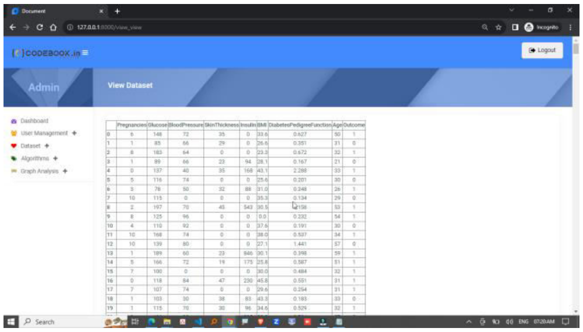
User Prediction Details



Diabetes Is Detected

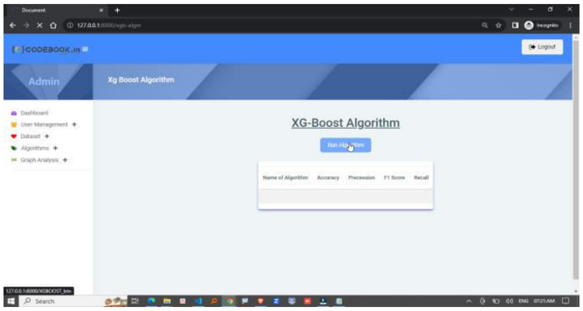
Accuracy: 82.47%
Recall: 80.4%
Precision: 76.65%
F1_Score: 80.4%

Prediction graph



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1
5	5	116	74	0	0	25.6	0.201	30	0
6	3	78	50	32	88	31.0	0.248	26	1
7	10	115	0	0	0	35.3	0.134	29	0
8	2	197	70	45	543	30.5	0.158	53	1
9	8	125	96	0	0	0.0	0.232	54	1
10	4	110	92	0	0	37.6	0.191	30	0
11	10	168	74	0	0	38.0	0.537	34	1
12	10	139	80	0	0	27.1	1.441	57	0
13	1	189	60	23	846	30.1	0.398	59	1
14	5	166	72	19	175	25.8	0.587	51	1
15	7	100	0	0	0	30.0	0.484	32	1
16	0	118	84	47	230	45.8	0.551	31	1
17	7	107	74	0	0	29.6	0.254	31	1
18	1	103	30	38	83	43.3	0.183	33	0
19	1	115	70	30	96	34.6	0.529	32	1

Data set



XG-Boost Algorithm

Name of Algorithm	Accuracy	Precession	F1 Score	Recall

Applying Algorithm

VII CONCLUSION

Multiple benchmark performance metrics like accuracy, precision and error in classification it will take into account to estimate the performance and efficiency of the proposed model.

The acquired results are validated by comparing them with the outcomes of traditional approaches employedin health sector domain and is observed to have shown promising performance. The inputs of several diabetic patients has been obtained from UCI laboratory which is further utilised to understand and locate patterns using ML algorithms like K Nearest Neighbors (KNN), ANN, XG Boosting and ada boosting, SVM . The simulated performance are compared and evaluated for performance and accuracy aspects. The proposed model produces a staggering outcome of around 98 percent, in comparison with other conventional approaches. Future research work has been on to enhance the security of the proposed system through adding intrusion detection based techniques.

REFERENCES

[1] A. Belle, R.. Thiagarajan, S. M. R. Soroushmehr, F. Navidi, D. A. Beard, and K. Najarian, ―medicalcare using with Big DataAnalytics ,‖ Hindawii Publ. Corp., vol. 2015, pp. 1–16, 2015..

[2] 2.. D. Yadav, A. Shinde, A. Nair, Y. Patil and S. Kanchan, ‖Enhancing the Data Security in Cloud ,‖2020 4th International Conference on Intelligent Computing and ControlSystems(ICICCS),2020,pp.753-757,,doi: 10.1109/ICICCS48265..2020.9121109.

[3] 3. Y. Ahmed, S. Naqvi and M. Josephs, ‖Cybersecrity Metrics for Protection of medicalcare IT Systems,..‖ 2019 13th International Symposium on Medical Information and Communication Technology (ISMICT), 2019, pp. 1-9, doi: 10.1109/ISMICT.2019.8744003...

[4] ―The big-data revolution in US health care: Accelerating value and innovation ― McKinseyamp;Company.[Online].Available:

https://www.mckinsey.com/industries/healthcaresystemsandservices/ our-insights/the-big-datarevolution-in-us- healthcare. [Accessed: 12-May2018]..

[5] 5. M. N. Bhuiyan, M. M. Rahman, M. M. Billah and D. Saha, ‖Internet of Things (IoT): A Review of Its En- abling Technologies in Healthcare Applications, Standard Protocols, Security, and Market Opportunities,‖ in IEEE Internet of Things Journal, vol. 8, no. 13, pp. 10474- 10498, 1 July1, 2021, doi: 10.1109/JIOT.2021.3062630. 2017.

[6] L. Zhou, S. Pan, J. Wang, and A. V. Vasilakos, ―big data in machine learning : Opportunitie and challenges,‖ Neurocomputing, vol. 237, pp.

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

[8] 8. Z. Ying,, W. Jiangg, X. Liu and S. Xu, ‖Implementing Security-Enhanced PHR System in the Cloud Using FAME,‖ 2019 IEEE Global Communications Conference (GLOBECOM), 2019, pp. 1-6, doi: 10.1109/GLOBECOM38437.2019.9014230.