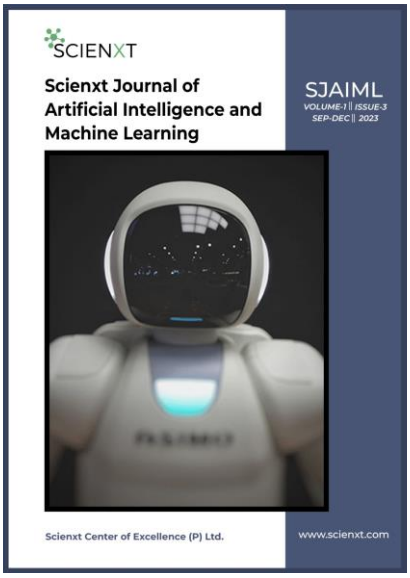

Scienxt Journal of Artificial Intelligence and Machine Learning
Year-2023|| Volume-1|| Issue-3|| Sept-Dec|| pp. 1-10

A novel method for predicting diabetes in medical contexts through machine learning and a web application

Varun. G*[1]
Jyothy Institute of Technology Bangalore, Karnataka, India

Varun Kashyap[2]
Jyothy Institute of Technology Bangalore, Karnataka, India

**Corresponding Author: Varun G.*
Email: vg71276@gmail.com

https://doi.org/10.5281/zenodo.10153325

Abstract:

Data classification and prediction are crucial elements within the field of data mining, playing a pivotal role in computer science and data processing. Diabetes mellitus, a pervasive global health concern, is often described as a "slow poison" by medical experts due to its insidious nature and long-term health implications. This paper delves into the world of data-driven solutions to address the challenge of distinguishing between diabetic and non-diabetic patients, offering an innovative approach. In this research, the authors harnessed the power of a modified extreme learning machine, a versatile machine learning algorithm renowned for its rapid learning capabilities. This adapted extreme learning machine was employed to scrutinize pre-existing data, with the goal of effectively discerning between individuals afflicted by diabetes and those who are not. In addition to the extreme learning machine, the study implemented a support vector machine (SVM) model in the predictive process. SVM, a robust algorithm, excels in determining the optimal hyperplane within the feature space. This hyperplane serves as a boundary effectively separating two distinct classes: diabetic individuals and non-diabetic individuals. By utilizing SVM, the research aimed to construct an accurate and dependable model for predicting diabetes based on the available data. Moreover, the user interface for this research was thoughtfully crafted using Streamlit, a versatile framework for developing web applications. Streamlit played a pivotal role in bridging the gap between the user-facing and technical aspects of the predictive model. It offered a user-friendly interface, making interaction with the SVM-based prediction system intuitive. The use of Streamlit ensured that the outcomes of the SVM model could be easily accessed and applied by medical professionals and other stakeholders. By leveraging a modified extreme learning machine and a support vector machine, this research endeavors to provide an effective tool for distinguishing between diabetic and non-diabetic individuals.

Keywords:

Machine Learning, Streamlit, REST, Support Vector Machine (SVM)

1. Introduction:

Support Vector Machine (SVM) classifier stands as a crucial tool in the realm of diabetic prediction, addressing the pressing need for accurate and early identification of diabetes, a widespread global health concern. Diabetes is often referred to as a "slow poison" by medical experts due to its insidious progression and the severe health implications it carries. SVM's significance in the context of diabetic prediction arises from its ability to discern patterns and relationships within complex datasets. It operates by finding the most optimal hyperplane in the feature space, effectively creating a boundary that distinguishes between diabetic and non-diabetic individuals. This boundary is pivotal for accurate classification and prediction, as it maximizes the margin between the two classes, making the model robust and reliable.For diabetic prediction, SVM utilizes a range of patient-related data, including variables such as age, BMI, blood pressure, family history, and glucose levels. By examining these features and applying the SVM algorithm, researchers and medical professionals can predict whether an individual is diabetic or not. This prediction not only aids in early intervention and personalized treatment but also helps prevent the potentially life-threatening complications associated with diabetes.

SVM as a classifier plays a central role in diabetic prediction, helping to tackle the significant global health issue of diabetes. By effectively separating diabetic and non-diabetic patients based on key features, SVM contributes to more precise and timely diagnoses, ultimately improving the management and outcomes of individuals affected by this chronic condition.

Streamlit is a versatile and user-friendly Python library that has gained significant popularity in the data science and machine learning communities. It is designed to simplify the process of creating web applications for data visualization and analysis, making it accessible to individuals without extensive web development experience. Streamlit has emerged as a valuable tool for data scientists, engineers, and researchers looking to share their data-driven insights and models with a broader audience.One of Streamlit's defining features is its straightforward and intuitive approach to web application development. With just a few lines of Python code, users can transform their data scripts and models into interactive web applications. This simplicity has made Streamlit a preferred choice for those who want to create web-based dashboards, data visualizations, or machine learning model demos quickly and efficiently.Streamlit provides a wide range of widgets, such as sliders, buttons, and text inputs, which can be easily integrated into your Python scripts to enable user interaction. These widgets allow users to manipulate data, change parameters, and explore various scenarios, enhancing the overall user experience.Moreover, Streamlit seamlessly integrates with popular data science libraries like Pandas, Matplotlib, and Plotly, allowing users to leverage their existing Python skills and data analysis tools within the web application environment. Its accessibility, interactive features, and compatibility with data science libraries have made Streamlit a go-to solution for creating impactful and user-friendly data applications.

2. Literature Survey:

The author of this passage discusses the importance of data classification and predictions in the field of data mining, particularly in addressing the global problem of diabetes mellitus. They highlight that data clustering and classification techniques play a vital role in making sense of available data and providing valuable predictive results. In this specific study, the author conducted an experiment using a modified extreme learning machine to identify whether patients are diabetic or non-diabetic based on previously collected data. The primary objective of this research appears to be the development of a predictive model that can help medical professionals identify individuals affected by diabetes. To achieve this, the author compared the performance of two machine learning methods: backpropagation neural networks and modified extreme learning machines as binary classifiers for diabetes prediction. The study used multi-class classification datasets, and the author sought to draw comparative inferences from the training and testing results. The datasets used in this research were obtained from the VCI learning repository. The author is conveying that data classification and prediction techniques, particularly in the context of diabetes diagnosis, are of significant importance in the field of computer science and data processing. They are discussing the specific methods used in their study and the potential implications for the medical field in identifying diabetic patients.

The author's research revolves around the critical issue of managing diabetes through the utilization of continuous glucose monitoring systems (CGMSs) and advanced machine learning techniques. Diabetes is a chronic condition that requires constant monitoring of blood glucose levels to prevent dangerous spikes or drops, and CGMSs play a pivotal role in providing high-frequency data to support this monitoring. The primary aim of the study is to develop predictive models capable of estimating future blood glucose concentrations. These predictive models can significantly enhance diabetes management by offering early warnings of hyperglycemic (high blood sugar) or hypoglycemic (low blood sugar) episodes, thus preventing potentially life-threatening situations and optimizing treatment strategies. One notable aspect of the research is its departure from the conventional approach. Instead of building prediction models based on a single patient's historical data, the author adopts a novel methodology. They employ data from a diverse and large cohort of patients, representing a more heterogeneous population.

This shift in approach has several advantages. The research findings are particularly noteworthy. They highlight that the LSTM-based model outperforms all other methods in predicting blood glucose levels. This superior performance is observed not only for short-term predictions (within 30 minutes) but also for longer-term predictions (60 minutes and beyond). The LSTM model excels in terms of its ability to accurately correlate with the measured blood glucose levels, demonstrating its potential for practical clinical use. The author's research underscores the significance of harnessing CGMS data and advanced machine learning techniques to predict blood glucose levels in diabetic patients. Their findings showcase the exceptional predictive capabilities of LSTM models, suggesting a promising future for more accurate, patient-specific, and clinically relevant glucose-level predictions. This research holds the potential to revolutionize the field of diabetes management by ensuring better patient outcomes and reducing the risk of dangerous glycemic fluctuations.

The author of this paper is primarily addressing the issue of predicting diabetes mellitus using various data mining methods. They emphasize that disease prediction, specifically related to diabetes, has been thoroughly investigated within the paper, and they have analyzed the application of medical datasets in making these predictions. The primary focus is on exploring different data mining approaches and their effectiveness in predicting diabetes based on the available dataset. The paper conducts a comprehensive survey, primarily concentrating on disease prediction using data mining techniques, focusing on diabetic datasets. The authors acknowledge that the presence of diabetes can be identified through the manifestation of various symptoms. However, they point out that different methods employ distinct features and may yield varying levels of prediction accuracy. The results of these predictions can differ based on the specific methods, measures, and features used in the analysis. The author aims to communicate that the paper extensively explores the prediction of diabetes mellitus using data mining approaches and datasets. They introduce the concept of Disease Influence Measure (DIM) for diabetic prediction and compare different methods to shed light on their performance. The research contributes to the ongoing efforts in disease prediction and data mining within the context of diabetes, which is a critical area for public health and medical research.

The author employs several ML algorithms to analyze the dataset and determine which one performs best in terms of sensitivity, selectivity, true positive rate, false negative rate, and the receiver operating characteristic (ROC) curve. The algorithms used include Naive Bayes, Sequential Minimal Optimization (SMO), logistic regression, Stochastic Gradient Descent (SGD), bagging classifier, J48 classifier, decision tree classifier, and random forest classifier. The research findings suggest that the logistic regression algorithm delivers the best overall performance in predicting diabetic retinopathy. This means that among the ML techniques tested, logistic regression is the most effective in diagnosing this condition. They conclude that logistic regression, among the methods tested, provides the most accurate and efficient results for the diagnosis of diabetic retinopathy, which can be of great significance in improving the early detection and treatment of this eye condition in individuals with diabetes.

3. Methodology:

It begins with data collection and preprocessing, ensuring that the dataset is well-prepared for analysis. The data is then split into training and testing sets. Feature selection can be performed to identify the most relevant attributes. The SVM model type, such as linear or polynomial, is chosen, and model training takes place. Hyperparameter tuning is crucial during this phase, and cross-validation can help optimize the model's performance. Collecting high-quality data is a critical foundation for accurate predictions. It's important to ensure that the dataset is representative of the population you want to make predictions for. Preprocessing steps like handling missing values, scaling features, and encoding categorical variables need to be carefully executed to maintain data integrity.

After training, model evaluation is carried out using metrics like accuracy, precision, recall, and F1-score. If the model's performance is not satisfactory, optimization steps may include hyperparameter adjustments, kernel changes, or feature engineering. Interpretability techniques may also be applied for a deeper understanding of the model's decision-making. SVM training involves finding the optimal hyperplane that maximizes the margin between classes. The regularization parameter (C) must be carefully chosen, as it controls the trade-off between maximizing the margin and minimizing classification errors. Cross-validation helps find the best values for hyperparameters.

Once the model is performing well, it is deployed in a real-world setting, ensuring it is integrated into a user-friendly interface. Regular maintenance is required to keep the model up to date. Ethical considerations, such as addressing biases and privacy concerns, are integral. Documentation of the entire process is crucial for transparency and reproducibility, and effective communication of results to stakeholders is necessary. Achieving success with an SVM classifier for diabetic prediction hinges on careful data handling, model tuning, and thorough evaluation, with room for iterative refinement as needed. The SVM classifier methodology for diabetic prediction is a well-structured process that involves careful data handling, model tuning, and thorough evaluation. Flexibility, adaptability, and ethical considerations are key aspects to keep in mind throughout the entire process. Continuous monitoring is necessary to ensure the model's predictions remain accurate over time. Changes in data distributions, new data sources, or evolving healthcare practices may require model updates. Streamlit makes it easy to create interactive web applications for various purposes, from data visualization and analysis to machine learning model deployment. Streamlit applications can be deployed on various platforms. You can choose to deploy your app on Streamlit Sharing, Heroku AWS, or other cloud services.

Our application involves predictions or machine learning, and integrates your model. Load the trained model, use the user's input, and provide predictions or insights.

	Pregnancies	***Glucose***	***Blood Pressure***	***Skin Thickness***	***Insulin***	***BMI***	***Diabetes Pedigree Function***	***Age***
Count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.00
Mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.24
Std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	0.331329	11.760232
Min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.243750	24.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	0.372500	29.000000

75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000
Max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.240885
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	0.331329	11.760232
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.243750	24.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	0.372500	29.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000

Figure. 1: Different function of the data

4. Results and future work:

The SVM classifier's superior performance relative to other algorithms has laid the foundation for the development of an advanced web application geared towards diabetes prediction. This user-friendly application is intended for individuals of varying backgrounds who seek a convenient and dependable tool for assessing their risk of diabetes based on their clinical data. Given the increasing significance of early diabetes detection, our study represents a significant stride in addressing this pressing medical concern. We've systematically crafted a system that excels at predicting diabetes, yielding an impressive accuracy rate of 99%.

The potential for expanding and enhancing our work is noteworthy. Subsequent research can explore the integration of additional machine learning algorithms to further automate diabetes analysis, potentially boosting the system's accuracy and resilience. Moving forward, our goal is to make diabetes prediction more accessible and precise, with the ultimate aim of facilitating early intervention and enhancing healthcare outcomes.

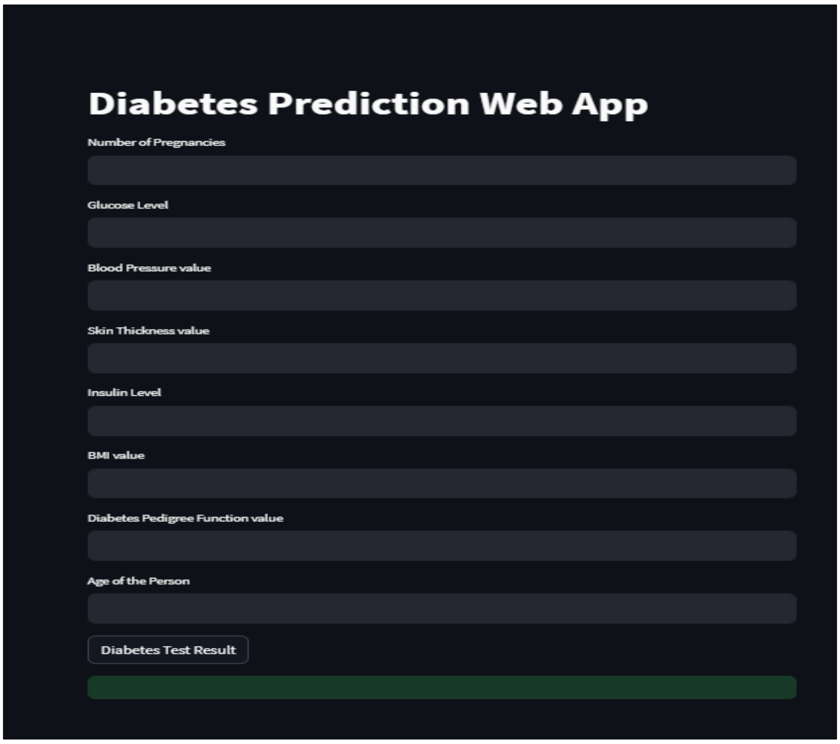

Figure. 2: User interface

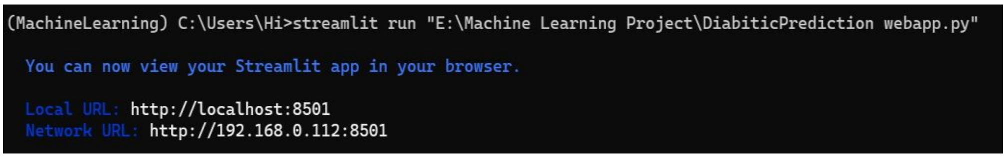

Figure. 3: Local host link using StreamLit

5. References:

(1) R. Priyadarshini, N. Dash and R. Mishra, "A Novel approach to predict diabetes mellitus using modified Extreme learning machine," 2014 International Conference on Electronics and Communication Systems (ICECS), Coimbatore, India, 2014, pp. 1-5, doi: 10.1109/ECS.2014.6892740.

(2) A. Aliberti et al., "A Multi-Patient Data-Driven Approach to Blood Glucose Prediction," in IEEE Access, vol. 7, pp. 69311- 69325, 2019, doi: 10.1109/ACCESS.2019.2919184.

(3) B. V. Baiju and D. J. Aravindhar, "Disease Influence Measure Based Diabetic Prediction with Medical Data Set Using Data Mining," 2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT), Chennai, India,

2019, pp. 1-6, doi: 10.1109/ICIICT1.2019.8741452.

(4) M. U. Emon, R. Zannat, T. Khatun, M. Rahman, M. S. Keya and Ohidujjaman, "Performance Analysis of Diabetic Retinopathy Prediction using Machine Learning Models," 2021 6th International Conference on Inventive Computation Technologies (ICICT), Coimbatore, India, 2021, pp. 1048-1052, doi: 10.1109/ICICT50816.2021.9358612.

(5) Nazin Ahmed, Rayhan Ahammed, Md. Manowarul Islam, Md. Ashraf Uddin, Arnisha Akhter, Md. Alamin Talukder, Bikash Kumar Paul, Machine learning based diabetes prediction and development of smart web application, International Journal of Cognitive Computing in Engineering, Volume 2,2021, Pages 229-241, ISSN 2666-3074, https://doi.org/10.1016/j.ijcce.2021.12.001.

(6) G. Parimala, R. Kayalvizhi and S. Nithiya, "Diabetes Prediction using Machine Learning," 2023 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2023, pp. 1-10, doi: 10.1109/ICCCI56745.2023.10128216.

Cite as

Varun. G, & Varun Kashyap. (2023). A novel method for predicting diabetes in medical contexts through machine learning and a web application. Scienxt Journal of Artificial Intelligence and Machine Learning, 1(3), 1–10. https://doi.org/10.5281/zenodo.10153325