Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7759534

DESIGN WEBSITE FOR DIABETES PREDICTION USING SUPERVISED MACHINE LEARNING

Sravanth Kumar Gaddibhukta
PG Research Scholar, Dept of CSE, Raghu Engineering College, Dakamarri, Visakhapatnam, Andhra Pradesh, India, sravanth1026@gmail.com

Dr. Panda B S
Professor, Dept of CSE, Raghu Engineering College, Dakamarri, Visakhapatnam, Andhra Pradesh, India, bspanda.cse@gmail.com

Karthik Vasanth Gaddibhukta
gvasanthkumar11@gmail.com

Srinija Vaddi
srinija.v122@gmail.com

ABSTRACT:
Many people are threatened with diabetes, making it one of the most serious diseases nowadays. Diabetes can be brought on by many factors, including advancing age, being overweight, not getting enough exercise, having a family history of the disease, leading an inactive lifestyle, eating improperly, having high blood pressure, etc. Diabetics are at increased risk for a wide range of health issues, including cardiovascular disease, kidney failure, stroke, vision and nerve problems, and more. Diabetic patients typically undergo a series of diagnostic examinations at the hospital before receiving therapy, as this is the standard procedure. The healthcare industry is a key user of big data analytics. The healthcare sector makes use of massive data sets. The study of massive datasets using big data analytics allows one to uncover previously unknown information and patterns, allowing one to draw conclusions and make predictions about the data. To improve diabetes classification, this research proposes a hybrid prediction model that combines the strengths of two Machine Learning (ML) models: the Support Vector Machine (SVM) and the Random Forest (RF). To evaluate the efficacy of the proposed hybrid model, the results of the model were compared with simple ML models, including SVM, RF, and Decision Tree (DT). The hybrid model with good accuracy is finalized and deployed in the web application. The website's layout makes it simple for visitors to navigate and use the interface to determine their health conditions.
KEYWORDS: Diabetes, Data Processing, Feature Extraction, Machine Learning, Accuracy, Website

INTRODUCTION:

Diabetes' global impact is expanding at an alarming rate, making it one of the most important issues confronting the healthcare sector today. As a result, according to the World Health Organization (WHO), globally diabetes is the 7th greatest reason for death in 2016 [1]. Diabetes-related complications kill an estimated 1.6 million individuals worldwide each year

[2]. The WHO reported in its first global study that the number of diabetics increased from 108 million in 1980 to 422 million in 2014 [3]. As part of World Diabetes Day on November 14, 2018, the World Health Organization (WHO) joined its global partners in highlighting the consequences of diabetes. According to WHO figures, one-third of adults are overweight, and the percentage is growing. Diabetes has been identified as the leading cause of cardiovascular disease, kidney disease, and stroke-related visual impairment. Diabetes is defined as a chronic disorder in which the body fails to correctly metabolize glucose (blood sugar; glucose is created by the food we eat), resulting in an abnormally high blood sugar content. Diabetes is characterized by resistance or insufficient production of insulin by the body. There was no known treatment for diabetes until recently, but early detection and education can help prevent problems. Diabetics face major dangers such as nerve damage, heart disease, kidney failure, and stroke. Hyperglycemia, caused by an excess of glucose in the blood, interferes with the regular functioning of several human organs. Early diagnoses are reliant on a physician's skill and experience, which may be incorrect or vulnerable. The healthcare business generates a massive amount of data, but it lacks the pattern recognition capabilities required to use that data to make informed decisions. Because of the inherent hazards of depending on a healthcare provider's subjective opinions and judgment for essential early-stage disease diagnosis [4]. Unknown patterns may exist, and they may have an impact on the outcomes we witness. As a result, patients are receiving inadequate care, necessitating the development of a more complex system for early disease identification, ideally one that employs automation to increase accuracy and efficiency. Data mining and ML techniques have advanced to the point where they can reliably and efficiently extract important insights from previously hidden data and find previously undetectable defects and hidden patterns. As the prevalence of diabetes rises, a plethora of data mining approaches has been created to find previously hidden trends in vast healthcare databases. Additionally, the data can be used for automated diabetes prediction and feature selection.

Algorithms trained using ML have the potential to significantly improve diabetes detection efforts, with applications ranging from predicting the most accurate diagnosis to automating a wide range of medical treatments. Among these are the use of various ML algorithms to help patients recognize the beginning of diabetes early. Diabetes prediction based on early symptoms can be made reasonably simple and inexpensive. This early projection can help people decide which treatment options to pursue and which experts to consult. As a result, in this study, we developed a system that consistently predicts diabetes based on health factors provided by the patient.

LITERATURE REVIEW:

The goal of the project is to [5] discover reliable ML models for recognizing diabetes cases from medical data. This article explains how to train ML algorithms with various datasets. To improve the quality of the models, we employed efficient label encoding and normalization pre-processing techniques. Furthermore, utilizing multiple feature selection approaches, we uncovered and prioritized a range of risk indicators. In-depth investigations on two independent datasets were conducted to assess the model's performance. The best ML method is chosen for further development. Flask, a web application platform written in Python, is used to implement such a concept in a website. This study suggests that utilizing an appropriate pre-processing strategy on medical data and ML-based categorization may improve the accuracy and effectiveness of diabetes prediction.

The study [6] evaluated and analyzed different ML models that can detect diabetes risk early on and aid in medical diagnosis. The project analyses real datasets, such as patient records acquired from a physician in the Bandipora district of India. ML model predictive abilities make them a valuable commodity in the healthcare market right now. ML model is being used by researchers to better forecast diseases and reduce expenditures. This research outlines a method developed by the author that uses ML methods to estimate the likelihood of getting diabetes in North Kashmir. Six ML classifiers have been used efficiently and RF achieves the best accuracy (98%) among classifiers with a well-balanced data set. Finally, this research helps to precisely predict the prevalence of diabetes and its future occurrence.

The work [7] is to give a thorough evaluation of diabetes diagnosis utilizing supervised and unsupervised ML approaches. This review includes papers published on diabetes diagnosis between 2018 and 2020. A DT-based algorithm was used to make accurate diabetes predictions. Unsupervised learning approaches like principle component analysis (PCA) and k-means clustering can also help with attribute selection and outlier detection from a huge dataset (K-Means). This study demonstrates that a mix of supervised and unsupervised ML algorithms employing K-Mean and SVM achieved high accuracy in diabetes diagnosis and evaluation.

The authors [8] present a model for diabetes prognosis that use a hybrid ML technique. The theoretical underpinnings of the overall framework include both SVM and Artificial Neural Networks (ANN). By analyzing the data, these models can establish if or not a patient suffers from diabetes. The dataset for this investigation is distributed 70:30 among training and testing data. Fuzzy logic is then employed to arrive at a final diagnosis of diabetes based on the results of such models serving as the membership function input to the fuzzy model. Models are combined and stored in the cloud for later use. This integrated approach may determine the person's diabetes condition based on their present medical profile. The suggested fused ML model has better prediction accuracy than other methods that have been previously reported.

The article [9] shows how to make a mobile app that uses Deep Learning (DL) to diagnose diabetes. The Sequential function of the TensorFlow platform was used to build the diabetes forecasting models. The model was converted to the 'tflite' type and used in the construction of a mobile app to determine whether a user is diabetic using the Android Studio IDE. The accuracy of the DL model was 93%, which was extremely impressive. Also, the app has instructions for how to use it and information about diabetes mellitus. The DL-based smartphone app that was just made is a big step forward in the field of early detection of diabetes mellitus. If the diagnosis is correct, a lifestyle modification may be achievable, allowing one to avoid a potentially disastrous conclusion.

To accurately diagnose and predict diabetes, as well as reduce its occurrence, a diabetes forecasting approach depends on a One-Dimensional Convolutional Neural Network (1DCNN) is suggested in the article [10]. To confirm its efficacy, we compared our proposed method to both the naive Bayes and the RF method. Several confusion matrix indicators were utilized to evaluate the three techniques. The experimental findings reveal that the suggested approach beats another model in precision and accuracy, and it is capable of accurately forecasting diabetes. The suggested 1DCNN approach has the potential to inspire new ways of data mining and analysis in the healthcare sector.

The study's [11] main objective is to forecast symptoms early, with the ultimate objective being to completely eradicate the illness. If the problem can be resolved whenever the illness remains in its initial stages, patients will have a significantly greater chance of getting better and not losing a lot of money. This study used a more severe ML and DL methodology to assess the model's effectiveness. When applied to a dataset, however, K-Nearest Neighbor and DT produce the maximum accuracy. Because the supplied data is well-suited for the KNN algorithm, it is the best option in this circumstance. A hospital in Bangladesh used a direct survey form to collect information from a wide range of patients, which was subsequently utilized in this study. The KNN algorithm is well suited for usage in medicine because of its mobility and a high degree of accuracy.

The research [12] uses an ML technique to predict the patient's blood sugar level based on actual data from a physical examination at a top-tier Chinese hospital. Several transformations are applied to the raw data to clean it up, and then, after a correlation is discovered between the feature variable and the target variable, some of the features are dropped. After the feature data has been given to XGBoost for feature selection, a risk prediction model could be used to classify high- and low-risk data. Experimentation yielded an improved model for predicting the chance of getting type II diabetes.

The study published in the journal [13] examines the effectiveness of 4 non-linear activation functions of the CNN approach used in blood glucose forecasting. Thirty patients are included in the simglucose UVA/Padova dataset utilized for experimental investigation in the simulation. The CNN approach also includes a three-class classification challenge in which the individual could be classified as euglycemic, hypoglycemia, and hyperglycemic. CNNs with various activation functions are assessed using statistical measures of their effectiveness. The results show that utilizing CNN with the ELU activation function yields the best accuracy.

METHODOLOGY:

The research tries to develop a website for diabetes prediction, which helps individuals to identify their health status without the doctor's need. To do the automatic diabetes prediction, the data will be collected from a standard website and processed. The collected data contain some unwanted information and unbalanced target data. The data processing is done to make the data into a clean and useful format. To remove unwanted data feature extraction is employed using a correlation plot. All the extracted features are given to the conventional ML models like SVM, RF, and DT and suggested hybrid SVM+RF for training. After completing the training process, all four ML models are tested, and the outcomes are evaluated using the metrics. This evaluation helps to identify the best model.

Data Acquisition

The dataset was created using the PIMA Indian dataset (PID) by NIDDK [14]. Using the PIMA dataset is important because most people around the world today share a similar lifestyle characterized by a higher intake of refined foods and a lower level of physical activity. As a result, the NIDDK has been undertaking the PID investigation, which is a cohort study, since 1965. Certain diagnostic features and measurements were incorporated into the dataset, enabling the early detection of chronic diseases, including diabetes. The collected data is shown in figure 1 and the following 8 parameters were shown to be most associated with diabetes prediction: the patient's age, number of pregnancies, body mass index, insulin levels, blood pressure, skin thickness, glucose levels, and Diabetes Pedigree Function.

	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

Figure 1- Diabetes PIMA Indian data

There were 768 incidences in PID, 268 were diabetic samples and 500 were non-diabetic samples. The resulting values are severely imbalanced. To resolve this issue, we can use the Synthetic Minority Oversampling Technique (SMOTE) to produce samples from the current data. Under no circumstances are copies made. Nitesh V. Chawla et al. reported the first SMOTE implementation in 2002 [15]. The essential assumption is that the sample's k-nearest neighbors are utilized to build a synthetic instance to represent the minority instances. In the specific situation, sample X is picked from the population using the k-nearest sampling strategy. The SMOTE technique then selects n samples at random and stores them in X_i. Finally, the following formula is used to generate a new sample X^'.

$$X' = X + rand(X_i - X), i = 1,2, \dots \dots, n$$

when rand is a random number selected at random from a uniform distribution over the specified interval (0, 1). ML approaches can be used to improve class imbalance by obtaining minority cases using SMOTE. Table 1 shows the data distribution before and after SMOTE.

Table 1. Data distribution before and after SMOTE

Diabetes Data Outcome	Before SMOTE	After SMOTE
0	500	500
1	268	500

Feature Extraction

Before applying the mining approach, the dataset must be pre-processed, which enables the elimination of redundant information and the counting of unstructured data by transformation. Theoretical ways to select important features differ depending on the problem. Employing Feature extraction is a critical intermediate stage in the mining process for streamlining the learning phase and improving performance without changing the fundamental architecture of data mining algorithms. The feature extraction strategy was implemented into the proposed method as a pre-processing phase to reduce the dimensionality of the dataset and speed up the entire computation procedure. Feature extraction is the process of extracting features from input data to enhance the performance of trained models [16]. This step in the overall structure eliminates unnecessary information, resulting in a decreased data dimension. The amount of time spent learning and inferring can be decreased. The feature extraction methods produce freshly generated features by combining and modifying the initial feature collection. The collected PIMA data contains 8 attributes, by using the correlation plot the 6 important attributes are retrieved from the data (eliminate blood pressure and skin thickness). The correlation plot used for feature extraction is given in figure 2. Figure 3 illustrates the extracted features from PID.

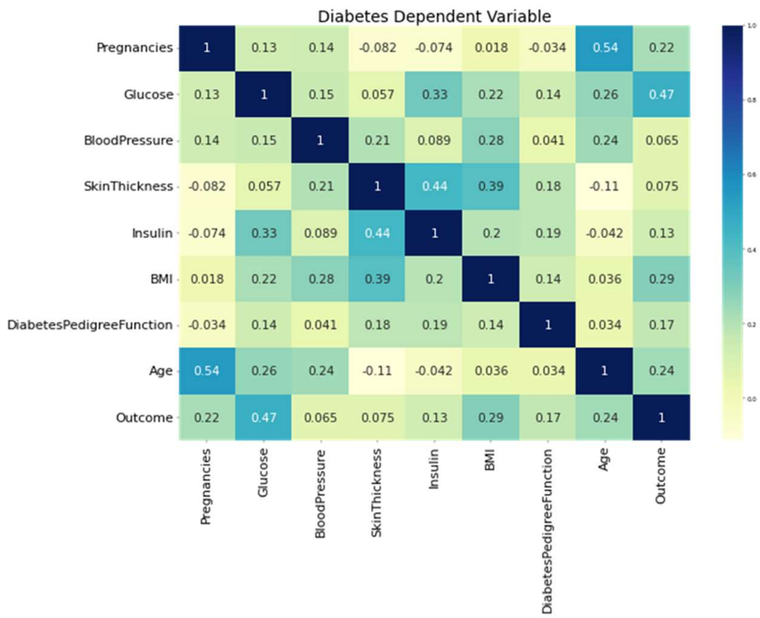

Figure 2- Correlation plot on diabetes data

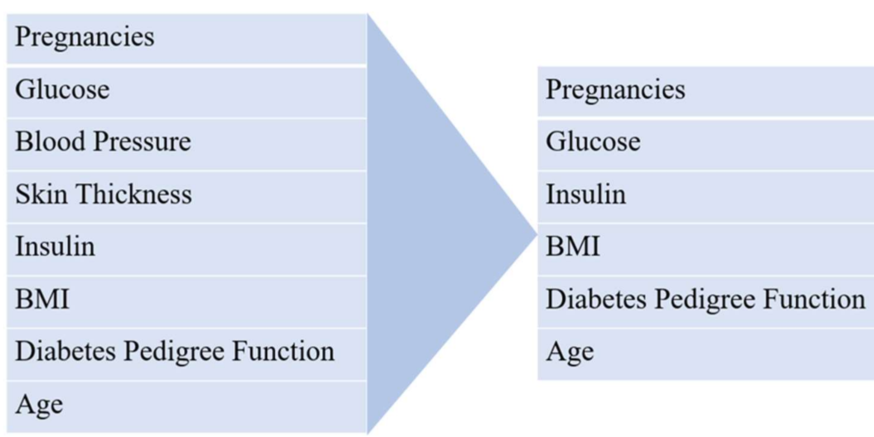

Figure 3- Extracted features using correlation plot.

ML Model

SVM algorithms are widely utilized for classification and regression in clinical applications [17]. While expanding the geometrical margin, SVM decreases the actual prediction error. A different name for SVM is Maximum Margin Classifiers. SVM is a broad process that relies on the assured risk constraints of optimization theory, known as the structural risk minimization principle. By applying the kernel method, SVM may efficiently execute non-linear categorization. Even though the feature space is uncertain, the kernel technique can be used to create the classifier. Each sample is represented as a point in space in an SVM model, and the categories are mapped onto nearby points to establish a sharp, as-wide-as-possible separation between them. Given a set of points from each class, an SVM, for example, can find the hyperplane with the maximum number of points from that class on the same plane. The optimal separating hyperplane minimizes the potential of misclassifying test dataset samples by maximizing the distance between two parallel hyperplanes.

A DT can divide a dataset into subsets based on a variety of criteria. A DT's root node serves as the beginning point for the tree's internal nodes, which are referred to as decision nodes. DTs can be used to address both regression and classification problems. It has an if-then-else structure that it follows. The leaves represent the result of categorizing various features with instances via the root node. Each node is chosen depending on the amount of data it contributes across all characteristics [18]. The logic behind a DT is detailed: (i) The features that serve as nodes are the tree's construction blocks. (ii) The input nodes with maximum information gain are utilized to choose features and forecast the output. (iii) By repeating the preceding steps, subtrees based on features not used in the root node are produced.

The term "random forest" refers to a collection of various DTs. Each tree predicts something based on a portion of the data, and the best one is chosen via a voting procedure. Averaging the output of each DT reduces overfitting even further. RF algorithm could be utilized for both regression and classification [19]. This is how RF functions: (i) A random sample is drawn from the provided dataset. (ii) For each sample, a DT is constructed, and the trees are utilized to create predictions. (iii) Every anticipated outcome is subject to vote. (iv) The most popular outcome is the one that occurs.

The fourth one is the design of the hybrid model. The goal of the hybrid method known as "ensemble prediction" is to reduce bias and variation while simultaneously improving predictions through the use of "boosting," "bagging," and "stacking." While Bayesian averaging is a novel ensemble technique, the following strategies have shown to be the most effective when attempting to synthesize results from multiple models: Boosting, wherein the forecast mistakes from earlier chain models are utilized to construct other models of a similar kind. Bagging is a technique for building numerous models from small samples of a single training sample. Stacking, wherein numerous models are constructed, with the supervisory model being the best possible amalgamation of the primary models' projections. This study addresses these problems and their impact on diabetes detection performance by carefully combining SVM and RF ML models to build an ensemble classifier via Bayesian averaging using efficient feature selection. Bayesian averaging involves making several forecasts for every piece of data. In this approach, we aggregate forecasts from different models and use the mean to construct our forecast. Predictions in regression and probability computations in classification tasks can both benefit from Bayesian averaging. Optimized feature selection is presented alongside a new ensemble learning approach to ensure resilience.

RESULTS AND DISCUSSION:

The PIMA India diabetes data were collected and processed for automatic diabetes prediction. The processed data is split into two, 80% for training and 20% for testing. The four ML models are trained using 80% data and tested using the remaining data. The outcome of test data is compared with the actual target value using the metrics. The highest accuracy was attained by SVM+RF (96.5%) followed by SVM (95%), RF (93.5%), and DT (89%). Next, the maximum recall value scored by SVM+RF (97.02%) followed by SVM (96.90%), RF (92.07%), and DT (90.90%), Then the precision score is analyzed and the values are presented in ascending order DT (87.37%), SVM (93.06), RF (94.89) and SVM+RF (96.07%). Finally, the F1-score is taken. SVM+RF produce gives the highest F1 value (96.55%), followed by SVM (94.94%), RF (93.46%), and DT (89.10%).

Table 2- Performance Comparison

MODEL	ACCURACY	RECALL	PRECISION	F1-SCORE
SVM	95	96.9072	93.0693	94.9495
DT	89	90.9091	87.3786	89.1089
RF	93.5	92.0792	94.898	93.4673
SVM+RF	96.5	97.0297	96.0784	96.5517

Table 2 is converted into a bar graph and it is shown in figure 4. For each metric individual bar graph is drawn. By analyzing all the bar graphs, it is found that the suggested SVM+RF model gives better results in all metrics.

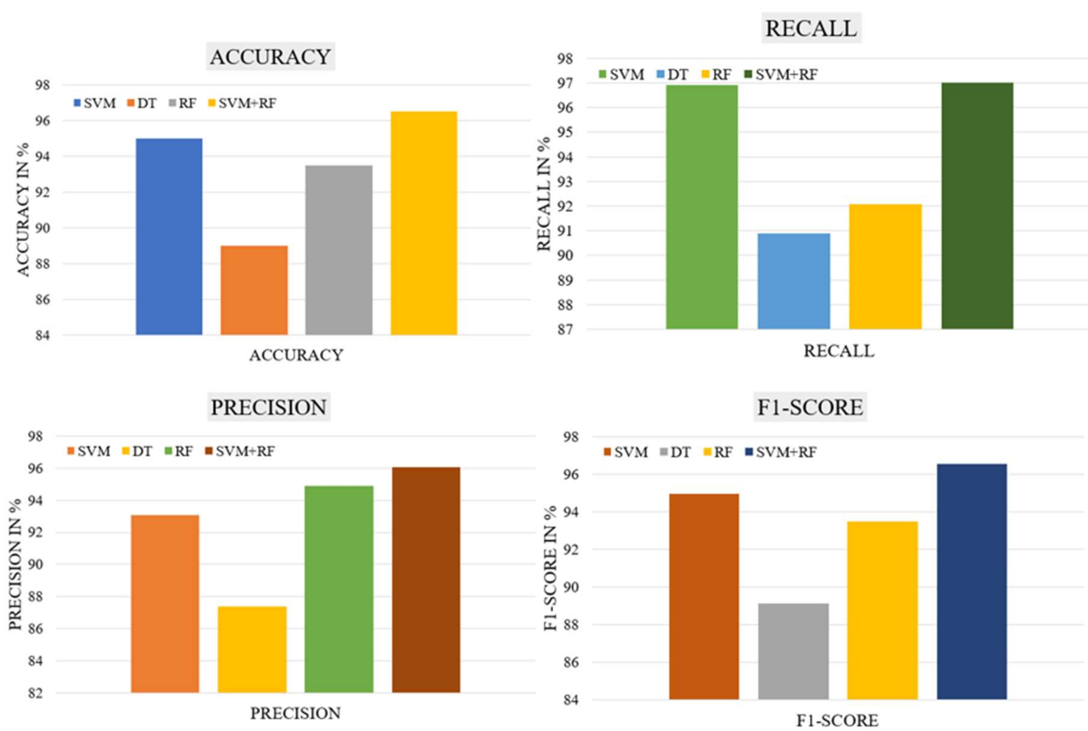

Figure 4- Performance analysis of ML models

After identifying the best model by comparing the metrics score, the best model is finalized and deployed on the website. A dedicated website was developed using the SVM+RF model and FLASK was used as a platform to develop one. On the website, different edit boxes were provided to make the user flexibility to enter their health parameters. After clicking on submit button, the user will get an accurate prediction of diabetes in a different dialogue box. The work of the designed website is given in figure 5.

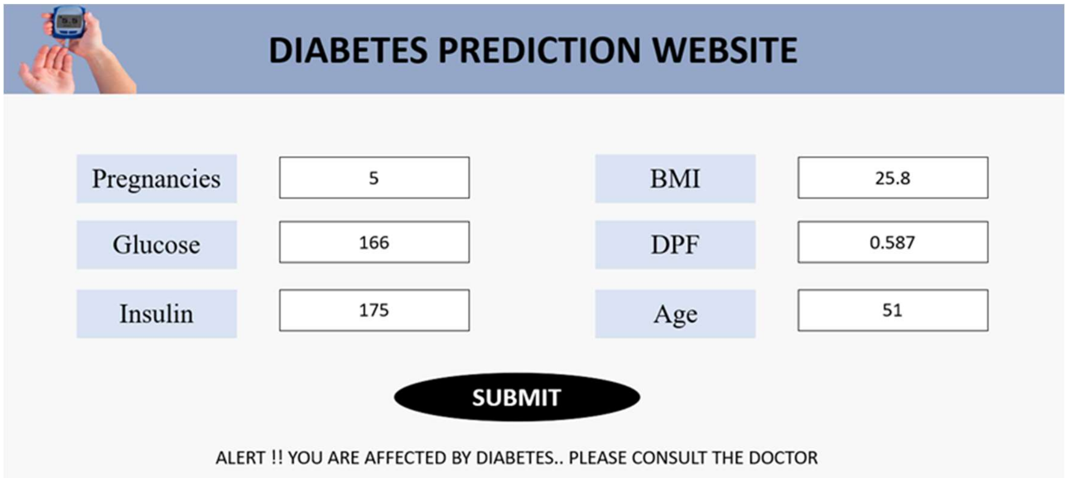

Figure 5- Working of the designed website for diabetes prediction.

CONCLUSION:

Diabetes is a long-term condition that makes the blood sugar level high. This long-term condition has been linked to early death. Diabetes can be diagnosed in a variety of ways, and its treatment can take several forms. Moreover, with the progress of technology, computer-aided disease diagnosis (CAD) is now feasible. CAD is a reliable, quick, and alternative approach that aids the physician's decision. Commonly utilized techniques for CAD includes ML classification strategies. The primary objective of this research was to develop an ML model for the early diagnosis of diabetes in patients. As a result, four ML classification techniques were utilized: SVM, RF, DT, and SVM+RF. The research was conducted using PID data obtained from the Kaggle website. The accuracy, precision, recall, and F1 score were used to rank the effectiveness of the ML approaches. The results show that the proposed hybrid SVM+RF technique is superior to others in accurately classifying diabetic disease. Finally, the simple website is intended to anticipate diabetes.

REFERENCE:

- World Health Organization (2021) https://www.who.int/news-room/fact-sheets/detail/diabetes, Accessed: 2021-04-20
- “Diabetes: Asia's 'silent killer'”, November 14, 2013”. Available at: www.bbc.com/news/world-asia-24740288.
- Emerging Risk Factors Collaboration; Sarwar N, Gao P, Seshasai SR, Gobin R, Kaptoge S, Di Angelantonio E, Ingelsson E, Lawlor DA, Selvin E, Stampfer M, Stehouwer CD, Lewington S, Pennells L, Thompson A, Sattar N, White IR, Ray KK, Danesh J. Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies. Lancet. 2010 Jun 26;375(9733):2215-22. doi: 10.1016/S0140-6736(10)60484-9. Erratum in: Lancet. 2010 Sep 18;376(9745):958. Hillage, H L [corrected to Hillege, H L]. PMID: 20609967; PMCID: PMC2904878.

- Naz H, Ahuja S. “Deep learning approach for diabetes prediction using PIMA Indian dataset”, J Diabetes Metab Disord, vol. 19, no. 1, pp. 391-403, 2020, doi: 10.1007/s40200-020-00520-5.
- Ahmed, Nazin & Ahammed, Rayhan & Islam, Manowarul & Uddin, Md Ashraf & Akhter, Arnisha & Talukder, Md. Alamin & Paul, Bikash Kumar. “Machine Learning based Diabetes Prediction and Development of Smart Web Application”, International Journal of Cognitive Computing in Engineering, vol. 2, 2021, doi: 10.1016/j.ijcce.2021.12.001.
- Shafi, Salliah & Selvam, Venkatesan & Ansari, Gufran & Ansari, Mohd Dilshad & Rahman, Md Habibur, “Prevalence and Early Prediction of Diabetes Using Machine Learning in North Kashmir: A Case Study of District Bandipora”, Computational Intelligence and Neuroscience, vol. 2022, pp. 1-12, 2022¸doi: 10.1155/2022/2789760.
- T. Chauhan, S. Rawat, S. Malik and P. Singh, "Supervised and Unsupervised Machine Learning based Review on Diabetes Care," 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2021, pp. 581-585, doi: 10.1109/ICACCS51430.2021.9442021.
- U. Ahmed et al., "Prediction of Diabetes Empowered With Fused Machine Learning," in IEEE Access, vol. 10, pp. 8529-8538, 2022, doi: 10.1109/ACCESS.2022.3142097.
- C. G. Estonilo and E. D. Festijo, "Development of Deep Learning-Based Mobile Application for Predicting Diabetes Mellitus," 2021 4th International Conference of Computer and Informatics Engineering (IC2IE), Depok, Indonesia, 2021, pp. 13-18, doi: 10.1109/IC2IE53219.2021.9649235.
- L. Xu, J. He and Y. Hu, "Early Diabetes Risk Prediction Based on Deep Learning Methods," 2021 4th International Conference on Pattern Recognition and Artificial Intelligence (PRAI), Yibin, China, 2021, pp. 282-286, doi: 10.1109/PRAI53619.2021.9551074.
- Saboor, A. U. Rehman, T. M. Ali, S. Javaid and A. Nawaz, "An Applied Artificial Intelligence Technique For Early Prediction of Diabetes Disease," 2022 Third International Conference on Latest trends in Electrical Engineering and Computing Technologies (INTELLECT), Karachi, Pakistan, 2022, pp. 1-6, doi: 10.1109/INTELLECT55495.2022.9969401.
- Z. Wei, W. Qiang and X. Yue, "Prediction of Type II Diabetes Risk Based on XGBoost and 1D-CNN," 2020 Chinese Automation Congress (CAC), Shanghai, China, 2020, pp. 5217-5222, doi: 10.1109/CAC51589.2020.9326662.
- S. Goel, S. Sharma and R. Tripathi, "Predicting Diabetes using CNN for Various Activation Functions: A Comparative Study," 2021 10th International Conference on System Modeling & Advancement in Research Trends (SMART), MORADABAD, India, 2021, pp. 665-669, doi: 10.1109/SMART52563.2021.9676280.
- https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database
- Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer, “SMOTE: synthetic minority over-sampling technique”, J. Artif. Int. Res, vol. 16, pp. 321–357
- Mutlag, Wamidh & Ali, Shaker & Mosad, Zahoor & Ghrabat, Bahaa Hussein, “Feature Extraction Methods: A Review”, Journal of Physics: Conference Series, vol. 1591, pp. 012028, 2020, doi: 10.1088/1742-6596/1591/1/012028.

- Christopher J.C. Burges. “A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery”, Springer, vol. 2, no. 2, pp.121-167, 1998.
- Iyer, A., S, J., Sumbaly, R., “Diagnosis of Diabetes Using Classification Mining Techniques”, International Journal of Data Mining & Knowledge Management Process, vol. 5, pp. 1–14, 2015, doi:10.5121/ijdkp.2015.5101, arXiv:1502.03774
- Schonlau, M., & Zou, R. Y, “The random forest algorithm for statistical learning”, The Stata Journal, vol. 20, no. 1, pp. 3–29, 2020¸ https://doi.org/10.1177/1536867X20909688