

Predictive Modelling for Diabetes and Insulin Dosage Using Machine Learning

Harshitha R¹ and Hemanth Kumar²

MCA Student, Department of MCA¹

Associate Professor, Department of MCA²

Jawaharlal Nehru New College of Engineering, Shivamogga, Karnataka, India

harshitha14601@gmail.com and hemanthkumar@jnnce.ac.in

Abstract: Now a days diabetes has become a chronic disease and managing this requires strict regular diet and workout to avoid various health issues and high blood glucose levels. To keep blood glucose at normal level in human body, diabetic patients have to be suggested with proper insulin dosage. It becomes difficult to predict the right amount of insulin to diabetes patients. To do this, Machine Learning (ML) method is used for identifying whether a person is suffering from diabetic and if he/she is suffering, right amount of insulin should be suggested to that patient. k-nearest neighbors (KNN) technique is employed to predict whether a patient is diabetic or not and Random Forest Regression technique is utilized for suggesting appropriate quantity of insulin dosage for the diabetic patient. Results are generated using the above-mentioned techniques.

Keywords: Diabetes prediction, Insulin dosage, k-nearest neighbors (KNN), Random Forest Regression, Pima Indian Diabetes Dataset, Machine Learning

I. INTRODUCTION

Diabetes mellitus is a universal continual sickness which effects the individual. This can be detected in those who do not have balanced glucose, metabolism in their body and lead to critical effect in health if not detected in early stages. It is dominant to detect the diabetes in the early stages. The early detection of diabetes helps in patient outcomes that will be enhanced with therapies which will reduce the likelihood of complications of diabetes and diagnosed on time. The identification of diabetes and prediction of suitable insulin dosages are critical factors in the treatment of the disease. Machine Learning, make it even possible to handle these issues. The main focus is to predict the insulin dosage who are high at risk. To do this, Machine Learning will help in identifying the possible issues in higher reliability and utilize the information. The k-nearest neighbors (KNN) and Random Forest Regressor is made used to employ the diabetes risk and insulin dosage by using dataset named Pima Indian Diabetes [16] which are obtained from medical field. Through the medical history of each patient which is present in dataset the female patients which is more than twenty-one-year-old including pregnancy, glucose, blood pressure, skin thickness, Insulin, Body mass index and age is considered. The main goal of the system is to identify the individual patient is diabetic or not based on selected attributes using the k-nearest neighbors (KNN) algorithm. When patient detected with diabetes, Random Forest Regressor is employed for suggesting right amount of insulin dosage. Random Forest Regressor calculates the correct amount of insulin required by diabetic person based on selected attributes. It can also be beneficial in achieving a significant value of the blood glucose levels as it reduces both hypo- and hyperglycemia which is also known as high and low blood glucose levels in patient. With the deployment of this model, prediction of risk in diabetes also adequate quantity of insulin dosage is enhanced using a Pima Indian Diabetes dataset.

II. LITERATURE SURVEY

Bilal M. Zahran et al. in [1] describes that managing the proper amount of the blood glucose will reduce the risk of people who are affected by diabetes and that will help in determining the best insulin dosage for the patients, this study utilized Artificial Neural Network (ANN) which includes various factors. In the study, a backpropagation-trained model was used, they trained the models, especially ANNs for predicting Type 1 and Type 2 diabetes.

Work in [2] by Malathi H et. al. presented that, which is having a Diabetes Advice System which gives Realtime Glucose Level Estimation that leverages ML algorithms such as RNN, XGBoost, and LSTM. It succeeds in adhering to data privacy and security but enhances accuracy of treatments.

Bhargava R et. al. describes the work in [3] that study on diabetes that to predict diabetes using deep learning methods. In this study utilized deep neural network which is achieved a high training accuracy of 98.07%.

Banibrata Paul et. al. in [4] proposed their work that, ANN is perfect tool to predict Diabetes mellitus, it recognized and got an accuracy of 77% to 100%. Thus, with early diagnosis of at least gestational diabetes, type 1 and also type 2 diabetes, these complications such as stroke and other diseases can be prevented.

A survey by S Kranthi Reddy et. al. in [5] presented their work that, diabetes also affects the heart and nerves. The Random Forest got accuracy of 78.4% and K-NN with 80.8% of accuracy at K=11 help in risk of diabetes and mitigation.

Madhumita Pal et. al. in [6] proposed that three Machine Learning based methods including K-NN, Random Forest, Linear SVM are employed in early prediction of the diabetes. It is come to know that Random Forest is best model with accuracy of 78.57 and AUC of 95.08 for the diabetes risk prediction which is more successful compared to Linear SVM and K-NN models.

The observation by Dr. K. Vijay Kumar et. al. in [7] discussed that by using Gradient Boosting Classifier to forecast the diabetes and insulin dosage is predicted by Linear Regression. In experimental results, and this work demonstrated that Gradient Boosting model accommodated effectively for high-performance of the results.

Contribution by Neelam Sai Kiran et. al. in [8] gives the overview of diabetes type 2, this study uses ANN model which predict right insulin dosage. Thus, the proposed ANN model, which used the backpropagation algorithm for training, might help in enhancing diabetes management through accurate insulin Prediction.

Work in [9] by Rouaa Alzoubi et. al. This study uses the machine learning and also uses the deep learning in the prediction of diabetes. Random Forest and k-nearest neighbors (KNN) have achieved the optimal accuracy of 98%.

Prateeksha Singh et. al. in [10] This study proposes in improving the diabetes prediction and classification accuracy of diabetes using the ML algorithm which is artificial neural network (ANN). Random Forest and KNN yield high accuracy for the prediction of diabetes. This work outperformed previous approaches with an accuracy of 99.13%.

Dr. M. Kalpana Chowdary et. al. in [11] In this piece of work author hereby presents an expert system which utilizes deep learning and also the machine learning to predict insulin dosage and diabetes and got 73% accuracy for ANN, 66% accuracy for Random Forest Regression Algorithm, & 85% accuracy for CNN Algorithm.

Srishti Mahajan et. al. in [12] presents an expert system using the approach of the machine learning i.e. Random Forest regression for insulin dosage calculation and deep learning i.e. Convolutional neural network (CNN) in Diabetes prediction. Random forest is applied and attains high accuracy level of up to 99.03%. Logistic regression classification can be implemented for good accuracy of 94.23%.

Sadia Afrin Shampa et. al. in [13] This study focuses on the machine learning (ML) model to predict diabetes within the populations from Bangladesh, India and Germany and used boosting techniques like the AdaBoost, CatBoost, Gradient Boost and XGBoost.

Na Hu et. al. in [14] This work addresses the early diagnosis of the Diabetes using various ML algorithms like KNN, decision tree, and Random Forest. Their work presented that the random forest model performed the best, with an accuracy of 0.84 and an F1 value of 0.77.

V Anirudh et. al. in [15] gives the overview of Diabetes mellitus may include kidney and heart illness. This work made use of the various ML algorithms in which gradient boosting classifier and logistic regression model aims at predicting diabetes and insulin dosage. These results present the potential and also for further development to increase the accuracy level.

III. METHODOLOGY

Machine Learning

Machine Learning (ML) is the branch of AI that enables an app, program or any device to learn from data and improve in the process. Problem solving structures may analyze data and information and make proper decisions based on it by applying AI and statistical patterns. ML model is a mathematical representation or algorithm which is trained on a

dataset to recognize patterns and to make predictions and perform various tasks. In Fig.1 shown the ML model. It has sub models

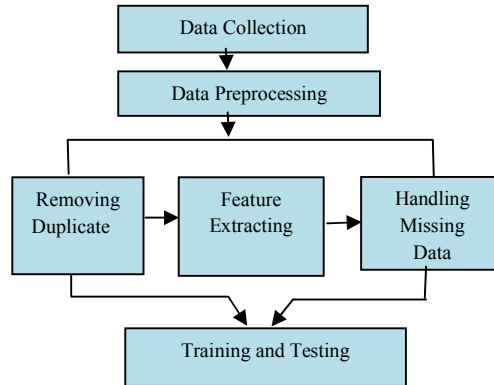


Fig. 1: Machine Learning model

Data Collection

This involves identifying the correct data to use. Collecting the correct type of data from surveys, experiments etc are undertaken.

Data Preprocessing

Data pre-processing includes steps such as data cleaning where duplicate entries are removed and normalization values are made standard and finally dataset is split into a training set and a test set.

Feature extraction

Feature extraction is the selection and modification of factors that put into formulation of the predictive model and other worthwhile influence.

Handling missing Data

Based on type of data to ensure quality of dataset is maintained, missing data in the dataset, which are either dropped or missing value is identified using statistical methods such as Median, Mean or Mode.

Training and Testing

The dataset is separated into training set and test set and training set is usually larger than the test set, for example 80% of the data which is made use for the training purpose while 20% made used for testing. Learning algorithms and their parameters are optimized on training data while the effectiveness and the ability of models to generalize is evaluated on the testing data.

k-nearest neighbors (KNN)

knn(KNN) is a ML algorithm which is used for classification and also for the regression tasks. This work uses knn algorithm for the prediction of diabetes. In Fig.2 shows working of k-nearest neighbors (KNN) algorithm. It works in the way that considers the distance of the nearest 'k' data points within the feature space. Also, an unseen data item gets assigned to the particular class by having the outcome of majority voting done on the nearest neighbors. KNN is a non-parametric method so, it does not require data to be from a given distribution. The algorithm employs that the distance calculations like the Euclidian distance to press out files with close similarity.

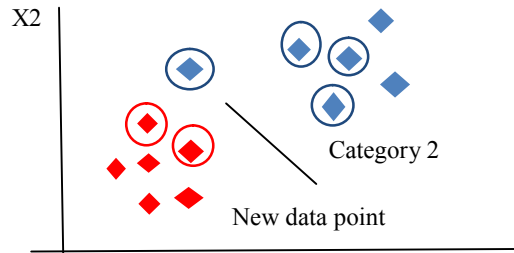


Fig.2: K-Nearest Neighbors

Random Forest Regression Algorithm

In Fig.3, It constructs number of decision trees, in training process with each of them being constructed based on different sets of attributes. By reducing the overfitting and capturing interactions between medical variables such as age, glucose, body mass index and skin thickness, this method is more accurate. Random Forest Regressor is part and which expands the use of algorithm which is used in the prediction of insulin dosage and the final output is obtained by combining each individual dosage estimates by each tree.

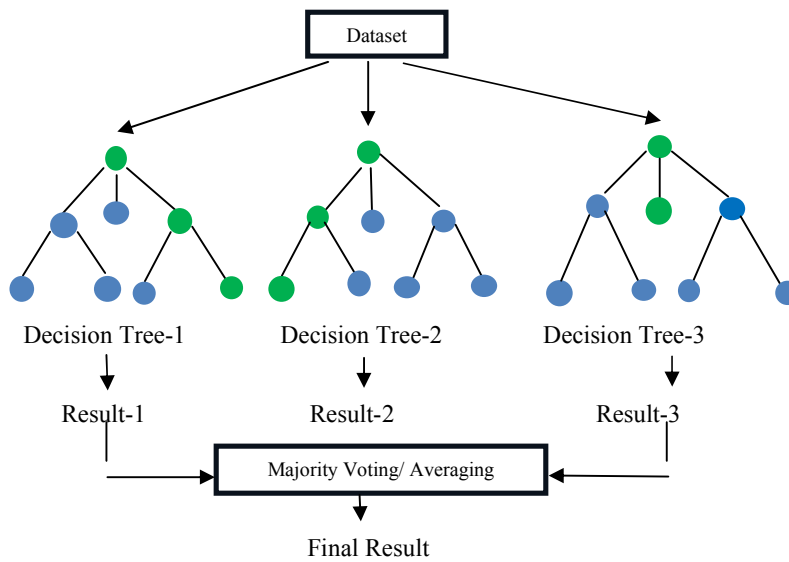


Fig.3: Random Forest regression

IV. PROPOSED MODEL

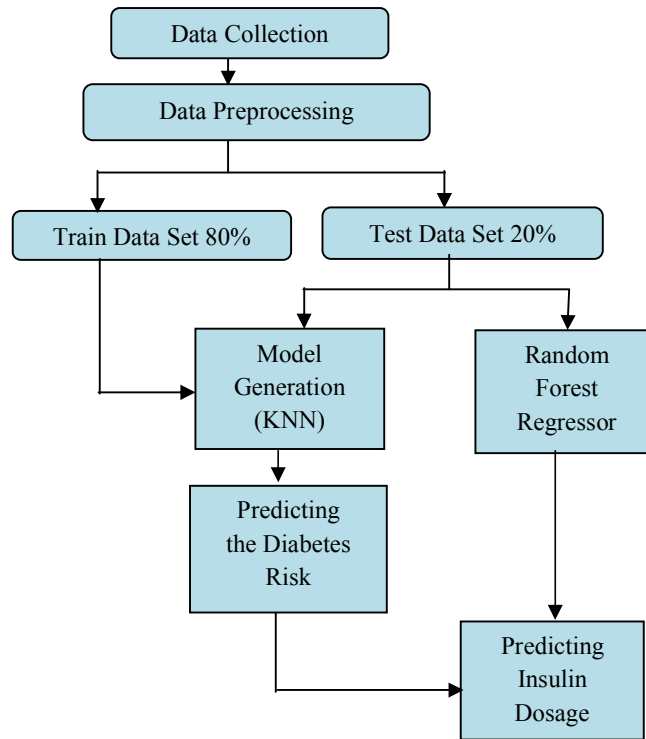


Fig.4: System model

1.Data Collection

Data is collected from various sources. The below table contains variables such as Pregnancies (Pre), Glucose level (Gluc), Blood pressure (BP), Skin thickness (ST) Insulin(Ins), Body Mass Index (BMI), Diabetes Pedigree Function (DPF) and Age. They are included in calculation of result.

TABLE 1: Dataset

PRE	GLUC	BP	ST	INS	BMI	DPF	AGE	OC
2	144	74	25	0	33.6	0.62	50	1
1	85	66	29	0	26.6	0.35	31	0
1	89	66	0	94	28.1	0.16	21	0
3	183	64	0	0	23.3	0.67	32	1
0	137	74	94	163	43.1	2.28	33	1

Data Preprocessing

Data preparation means process of data conditioning where complicated raw data is transformed into a format that can easily use to make analysis. It involves transforming certain features such as converting numerical features into smaller range of values, that is scaling or normalizing, transforming categorical features into numerical form, that is encoding and dealing with any missing value.

Model Generation

This contains only a 20% testing data and 80% of dataset is training data. For pre-processed datasets that is after feature extraction and training data are made used for training the model with the correct parameters. Later, model is being tested to determine how well it performs with test data.

Diabetes Risk Prediction

k-nearest neighbors(KNN) is employed for prediction of diabetes risk and the diabetes risk is provided in percentage value.

Insulin dosage prediction

Random Forest Regressor is utilized for predicting insulin dosage. knn model presents the emergence of diabetes risk. According to Random Forest regressor, if the patient has diabetes, then the insulin dosage is predicted.

V. RESULTS

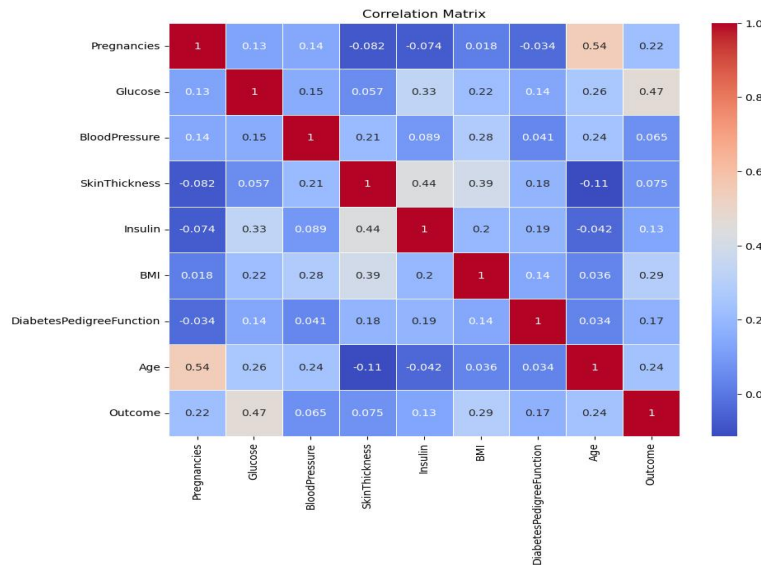


Fig.5: Correlation Matrix

The dataset availability and correlations between features are depicted in the heatmap correlation matrix in Fig.5. Darker color represent stronger correlations between two variables and each cell in matrix shows the correlation coefficient, indicating the strength and direction of the relationship between two variables. By employing this visualization, coefficients correlation between variables for the diabetes prediction can be determined.

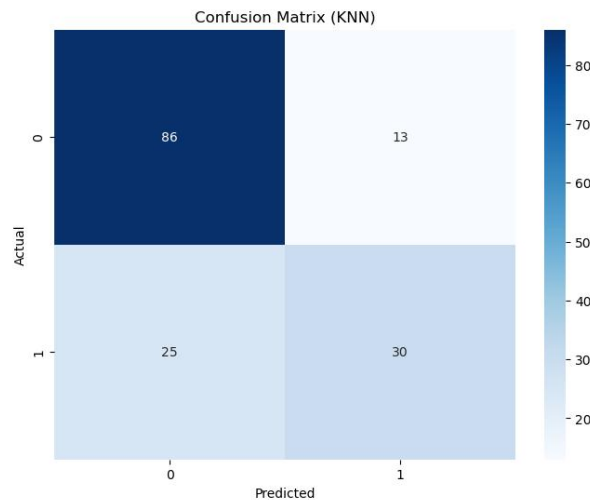


Fig.6: Confusion Matrix Heatmap

The results of our model are based on dataset which are presented as confusion matrix is shown in the Fig.6. In this each cell denotes the exact number of predictions made by system regarding the actual and anticipated state of condition of diabetes. The off-diagonal elements apply to mistaken identification, while the diagonal elements represent correct Prediction. This matrix which provides the accuracy of model in categorizing diabetes by defining the detailed accuracy mentioned in the Eq (1) [3].

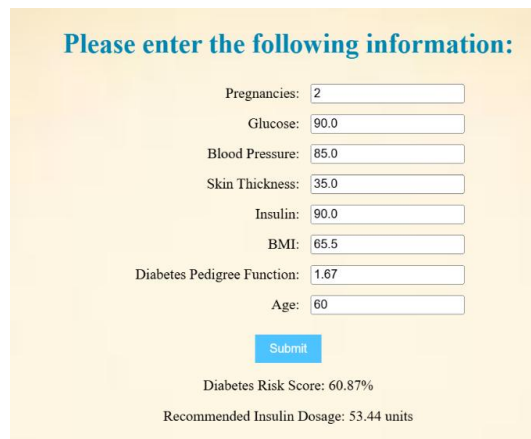
$$\begin{aligned}
 \text{Accuracy} &= \frac{TP+TN}{TP+TN+FP+FN} \\
 &= \frac{30+86}{30+86+13+25} \quad (1)
 \end{aligned}$$

True Positives (TP): Which Represents number of instances in which positive class is predicted by the model. (bottom-right cell)

True Negatives (TN): Which represents number of instances in which negative class is predicted by the model. (top-left cell)

False Positives (FN): Represents number of instances in which positive class is incorrectly predicted by the model. (top-right cell)

False Negatives (FN): Represents number of instances in which negative class is incorrectly predicted by the model. (bottom-left cell)



Please enter the following information:

Pregnancies:

Glucose:

Blood Pressure:

Skin Thickness:

Insulin:

BMI:

Diabetes Pedigree Function:

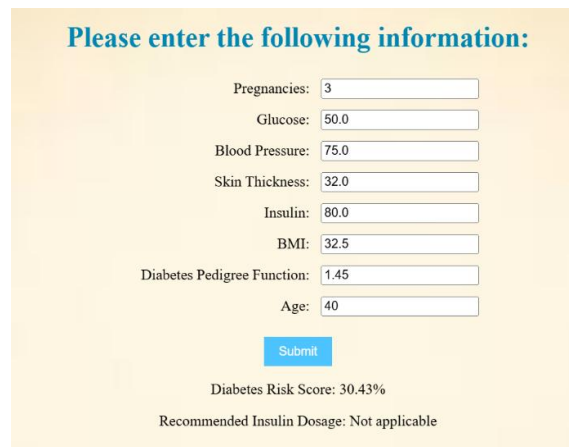
Age:

Diabetes Risk Score: 60.87%

Recommended Insulin Dosage: 53.44 units

Fig.7: Patient data input screen with results

Fig.7It shows the user interface for the work.In which, values are entered for different variables and risk of diabetes is provided in % value



Please enter the following information:

Pregnancies:

Glucose:

Blood Pressure:

Skin Thickness:

Insulin:

BMI:

Diabetes Pedigree Function:

Age:

Diabetes Risk Score: 30.43%

Recommended Insulin Dosage: Not applicable

Fig.8: Patient data input screen with results

It shows the user interface for the work. In which, values are entered for different variables and the risk of diabetes is provided in % value. In Fig.8, the diabetes score is lesser than 50%. The application concludes that the patient is not having diabetes. In this particular case, model does not recommend an insulin dose.

VI. CONCLUSION

This work establishes how the diabetes may be treated using machine learning where the Pima Indian Dataset is employed in our work. This methodology shows how the algorithms can provide diabetes patients with the early prediction of diabetes and the recommended insulin dosage. In this work Django is used as the web framework in our approach to make data processing settle and include user input. Altogether, this work discusses the outstanding role of ML in forecasting diabetes risk and insulin dosage and also the prevention and management of diabetes.

REFERENCES

- [1]. Bilal M. Zahran, "A Neural Network Model for Predicting Insulin Dosage for Diabetic Patients". International Journal of Computer Science and Information Security (IJCSIS), Vol. 14, No. 6, 2016.
- [2]. Malathi. H, D. C. Yadav and T. V. G, "Predictive Analytics and Machine Learning for Personalized Diabetes Management in Real-Time", *IEEE International Conference on ICT in Business Industry & Government (ICTBIG)*, 2023.
- [3]. R. Bhargava and J. Dinesh, "Deep Learning based System Design for Diabetes Prediction", *International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON)*, 2021, pp.1-5.
- [4]. B. Paul and B. Karn, "Diabetes Mellitus Prediction using Hybrid Artificial Neural Network", *IEEE Bombay Section Signature Conference (IBSSC)*, 2021, pp.1-5.
- [5]. S. K. Reddy, T. Krishnaveni, G. Nikitha and E. Vijaykanth, "Diabetes Prediction Using Different Machine Learning Algorithms", *Third International Conference on Inventive Research in Computing Applications (ICIRCA)*, 2021, pp-1261-1265.
- [6]. M. Pal, S. Parija and G. Panda, "Improved Prediction of Diabetes Mellitus using Machine Learning Based Approach", *2nd International Conference on Range Technology (ICORT)*, 2021, pp.1-6.
- [7]. Dr. K. Vijay Kumar, K. Mounika Sai Sadhvi, M. Bhargavi, K. Sri Varshini, K. Sirisha, "An Expertise System for Insulin Dosage Prediction using Machine Learning Techniques", International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering, Vol. 10, Issue 6, 2022.
- [8]. Neelam Sai Kiran, M. Keerthana, P. Bhavani, D. Revanth, P. Siva prasad, "Insulin dosage prediction system for diabetic patients", South Asian Journal of Engineering and Technology, 2022.
- [9]. Rouaa Alzoubi and S. Harous, "Machine Learning Algorithms for Early Prediction of Diabetes a Mini-Review", *International Conference on Electrical and Computing Technologies and Applications (ICECTA)*, 2022.
- [10]. P. Singh, S. Silakari and S. Agrawal, "An Efficient Deep Learning Technique for Diabetes Classification and Prediction Based on Indian Diabetes Dataset", *3rd International Conference on Technological Advancements in Computational Sciences (ICTACS)*, 2023.
- [11]. Dr. M. Kalpana. Chowdary, C. Ganesh, P. M. P. Raj, M. G. Kumar, M. Sridhar and S. Sandhya, "An Expert System for Insulin Dosage Prediction using Machine Learning & Deep Learning Algorithms," *2023 8th International Conference on Communication and Electronics Systems (ICCES)*, 2023, pp. 1291-1297.
- [12]. S. Mahajan, P. K. Sarangi, A. K. Sahoo and M. Rohra, "Diabetes Mellitus Prediction using Supervised Machine Learning Techniques", *International Conference on Advancement in Computation & Computer Technologies (ICACCT)*, 2023, pp. 587-592.
- [13]. S. A. Shampa, M. S. Islam and A. Nesa, "Machine Learning-based Diabetes Prediction: A Cross-Country Perspective", *International Conference on Next-Generation Computing, IoT and Machine Learning (NCIM)*, 2023.

- [14]. N. Hu and J. Gao, "Research on Diabetes Prediction Model Based on Machine Learning Algorithms" *International Conference on Computers, Information Processing and Advanced Education (CIPAE)*, 2023, pp. 200-203.
- [15]. V Anirudh, G Siri, Puli Shashank, T Rahul Vardhan, Veera Bhadra Rao, "Prediction of Diabetes and Insulin Dosage", *International Journal of Advances in Engineering and Management (IJAEM)* Vol 5, No.4, 2023, pp: 877-883.
- [16]. <https://www.kaggle.com/datasets>