IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A MACHINE LEARNING MODEL FORECASTING INSULIN DOSAGE

[1]Tanusha Budi, [2]Mondru Sion Kumari, [3]Tatakuntla Jahnavi, [4]Undi Meghana, [5]Usarti Mutyalamma, [6]Usarti Bindu Bhargavi

[1]Student, [2]Project Guide, [3]Student, [4]Student, [5]Student, [6]Student,
[1]Department Of Computer Science and Systems Engineering,
[1]Andhra University College Of Engineering For Women, Visakhapatnam, India

Abstract: Diabetes Mellitus, a chronic metabolic disorder, requires careful management of blood glucose levels (BGLs) to mitigate the risk of serious long-term complications. Despite the importance of traditional preventive measures like maintaining a healthy diet and regular exercise, many diabetic patients struggle to control their BGLs effectively. Proper insulin dosage plays a crucial role in managing this condition. Our project aims to leverage machine learning techniques to aid in diabetes prediction and insulin dosage estimation. We utilize the PIMA diabetes dataset and the UCI insulin dosage dataset for training our models. The Gradient Boosting Classifier is employed to predict the presence of diabetes, while the Linear Regression algorithm is used to estimate insulin dosage for diagnosed diabetic patients. Following training, we will assess the models' performance on a test dataset lacking class labels. The Gradient Boosting Classifier will identify diabetes cases, and for those diagnosed, the Linear Regression model will predict the appropriate insulin dosage. By integrating these predictive models, we aim to contribute to improved diabetes management strategies.

Index Terms **-** Machine Learning, Gradient Boosting Classifier, Random Forest, Python programming, PIMA Diabetes dataset, UCI (*University of California), Irvine* Insulin dosage dataset.

I. INTRODUCTION

The ability to predict glucose concentrations is crucial for timely patient intervention, especially in critical scenarios like hypoglycemia. Recent research has focused on employing advanced data-driven techniques to develop accurate models of glucose metabolism. Given the complex, non-linear, and patient-specific relationship between input variables (e.g., medication, diet, physical activity, stress) and glucose levels, non-linear regression models such as artificial neural networks, support vector regression, and Gaussian processes are being explored. As diabetes becomes increasingly prevalent in modern society, rapid and accurate diagnosis and analysis are paramount. Diagnostic criteria in medicine typically involve fasting blood glucose, glucose tolerance, and random blood glucose levels. Early diagnosis facilitates better management. Machine learning holds promise in leveraging daily physical examination data to provide preliminary assessments of diabetes mellitus, offering valuable insights for healthcare professionals. Key challenges in employing machine learning methods include selecting relevant features and choosing appropriate classifiers.

II. LITERATURE SURVEY

2.1 Utilizing Data Mining Techniques to Predict Diabetes Mellitus

Diabetes, characterized by elevated blood sugar levels, is a chronic disease with widespread implications for human health. Automated information systems employing various classifiers have been developed to anticipate and diagnose diabetes, leveraging data mining approaches. Early prediction is vital for effective disease management and can potentially save lives. Selecting appropriate classifiers is crucial for enhancing the accuracy and efficiency of these systems, particularly given the rising prevalence of diabetes worldwide. Many individuals are unaware of their diabetes risk factors before diagnosis, underscoring the importance of early detection. This study focuses on early diabetes prediction using data mining techniques, utilizing a dataset comprising 768 instances from the PIMA Indian Diabetes Dataset. Five predictive models were developed using nine input variables and one output variable. These models were evaluated based on accuracy, precision, sensitivity, specificity, and F1 Score measures. The aim is to compare the performance of Naïve Bayes, Linear Regression, Artificial Neural Networks (ANNs), C5.0 Decision Tree, and Support Vector Machine (SVM) models in predicting diabetes using common risk factors. Results indicate that the decision tree model (C5.0) achieved the highest classification accuracy, followed by the linear regression model, Naïve Bayes, ANN, and SVM, which exhibited the lowest accuracy. This study underscores the potential of data mining techniques in facilitating early diabetes prediction and highlights the comparative effectiveness of various predictive models.

2.2 Examining Different Data Mining Approaches for Diabetes Mellitus Prediction

Utilizing data mining techniques aids in diagnosing patients' diseases, including chronic conditions like Diabetes Mellitus, which can affect multiple organs. Early prediction is crucial, as it can potentially save lives and enable better disease management. This paper delves into early diabetes prediction using diverse data mining methods. The dataset comprises 768 instances from the PIMA Indian Dataset, enabling an assessment of the accuracy of these techniques. The analysis reveals that the Modified J48 Classifier exhibits the highest accuracy compared to other methods.

2.3 Comparing Data Mining Techniques for Predicting Diabetes Mellitus

Diabetes, stemming from elevated blood sugar levels, is a chronic ailment affecting numerous individuals worldwide. Employing various automated algorithms aids in the anticipation and diagnosis of diabetes. Utilizing data mining methods can assist in diagnosing patients' diseases, potentially saving lives by enabling preventive measures before the disease manifests. The selection of appropriate classification methods enhances the accuracy of the system, crucial given the rising prevalence of diabetes. Despite this, many diabetics remain unaware of their risk factors until diagnosis. This study develops five predictive models using nine input variables and one output variable to compare the performance of Naive Bayes, Decision Tree, SVM, KNN, and ANN models for predicting diabetes mellitus. Diabetes, characterized by hyperglycemia, affects a significant portion of the global population and can lead to long-term damage to various organs. The disease results from defects in insulin secretion or action, causing chronic hyperglycemia and subsequent organ dysfunction. Insulin deficiency, stemming from improper insulin secretion, is a primary contributor to hyperglycemia, highlighting the importance of insulin dosage. Predicting glucose concentrations is crucial for appropriate patient management, particularly in scenarios such as hypoglycemia.

III. PROPOSED SYSTEM

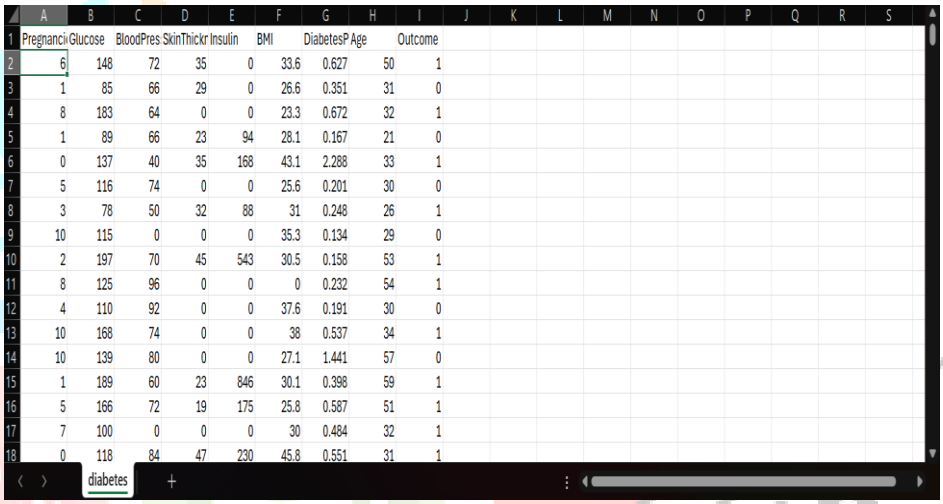
3.1 Methodology

In this project, we adopt a two-step approach for predicting diabetes and estimating insulin dosage in diagnosed diabetic patients. Initially, we employ a Gradient Boosting Classifier to predict the presence of diabetes using the PIMA diabetes dataset. This dataset is specifically designed for diagnostic prediction based on certain diagnostic measurements. Concurrently, we utilize a Linear Regression algorithm to estimate insulin dosage using the UCI insulin dosage dataset. The UCI Machine Learning Repository provides a comprehensive collection of databases and data generators essential for empirical analysis in the machine learning domain. Following the training phase with the mentioned

datasets, we proceed to test our models by uploading a dataset without class labels. The Gradient Boosting Classifier then predicts the presence of diabetes, while the Linear Regression model estimates insulin dosage for patients identified as diabetic by the Gradient Boosting Classifier. This method enables us to leverage both datasets effectively for accurate diabetes prediction and insulin dosage estimation.

3.2 Dataset

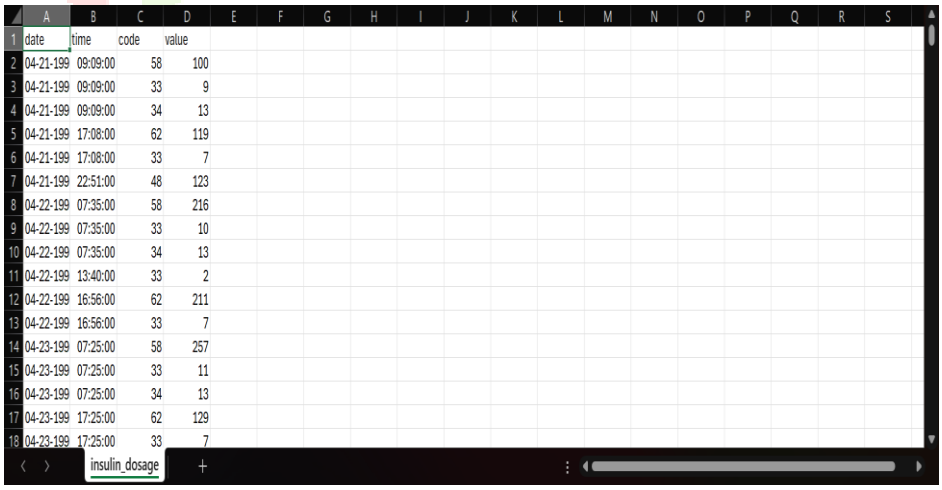
The dataset utilized in this study is the PIMA Indian Diabetes Dataset, sourced from the National Institute of Diabetes and Digestive and Kidney Diseases. It encompasses data from 768 women residing near Phoenix, Arizona, USA. Among the participants, 258 tested positive for diabetes, while 500 tested negative. The dataset comprises one target variable, which is the presence or absence of diabetes, along with eight attributes: pregnancies, Oral Glucose Tolerance Test (OGTT) results, blood pressure, skin thickness, insulin levels, Body Mass Index (BMI), age, and pedigree diabetes function. The Pima population has been subject to periodic study by the National Institute of Diabetes and Digestive and Kidney Diseases since 1965, reflecting the longitudinal nature of the dataset. Given the multifactorial nature of Type 2 Diabetes Mellitus (T2DM), the dataset includes information on attributes that may influence the onset of diabetes and its subsequent complications, acknowledging the interplay between genetic predisposition and environmental factors.



	A	B	C	D	E	F	G	H	I
1	Pregnanci	Glucose	BloodPres	SkinThickr	Insulin	BMI	DiabetesP	Age	Outcome
2	6	148	72	35	0	33.6	0.627	50	1
3	1	85	66	29	0	26.6	0.351	31	0
4	8	183	64	0	0	23.3	0.672	32	1
5	1	89	66	23	94	28.1	0.167	21	0
6	0	137	40	35	168	43.1	2.288	33	1
7	5	116	74	0	0	25.6	0.201	30	0
8	3	78	50	32	88	31	0.248	26	1
9	10	115	0	0	0	35.3	0.134	29	0
10	2	197	70	45	543	30.5	0.158	53	1
11	8	125	96	0	0	0	0.232	54	1
12	4	110	92	0	0	37.6	0.191	30	0
13	10	168	74	0	0	38	0.537	34	1
14	10	139	80	0	0	27.1	1.441	57	0
15	1	189	60	23	846	30.1	0.398	59	1
16	5	166	72	19	175	25.8	0.587	51	1
17	7	100	0	0	0	30	0.484	32	1
18	0	118	84	47	230	45.8	0.551	31	1

diabetes

Figure 1: Diabetes Dataset



	A	B	C	D
1	date	time	code	value
2	04-21-199	09:09:00	58	100
3	04-21-199	09:09:00	33	9
4	04-21-199	09:09:00	34	13
5	04-21-199	17:08:00	62	119
6	04-21-199	17:08:00	33	7
7	04-21-199	22:51:00	48	123
8	04-22-199	07:35:00	58	216
9	04-22-199	07:35:00	33	10
10	04-22-199	07:35:00	34	13
11	04-22-199	13:40:00	33	2
12	04-22-199	16:56:00	62	211
13	04-22-199	16:56:00	33	7
14	04-23-199	07:25:00	58	257
15	04-23-199	07:25:00	33	11
16	04-23-199	07:25:00	34	13
17	04-23-199	17:25:00	62	129
18	04-23-199	17:25:00	33	7

insulin_dosage

Figure 2: Insulin Dosage Dataset

IV. ALGORITHMS

4.1 Gradient Boosting

Gradient boosting is a versatile machine-learning technique utilized for regression and classification tasks, among others. It constructs a prediction model by combining an ensemble of weak prediction models, commonly decision trees. Specifically, when decision trees. XG Boost stands out as an optimized distributed gradient boosting library engineered to offer high efficiency, flexibility, and portability. It implements various machine learning algorithms within the Gradient Boosting framework.

The process of gradient boosting entails three fundamental steps:

1. Generating a classification dataset using the make classification method.
2. Constructing a Gradient Boosting Classifier.
3. Employing the classification model for making predictions.

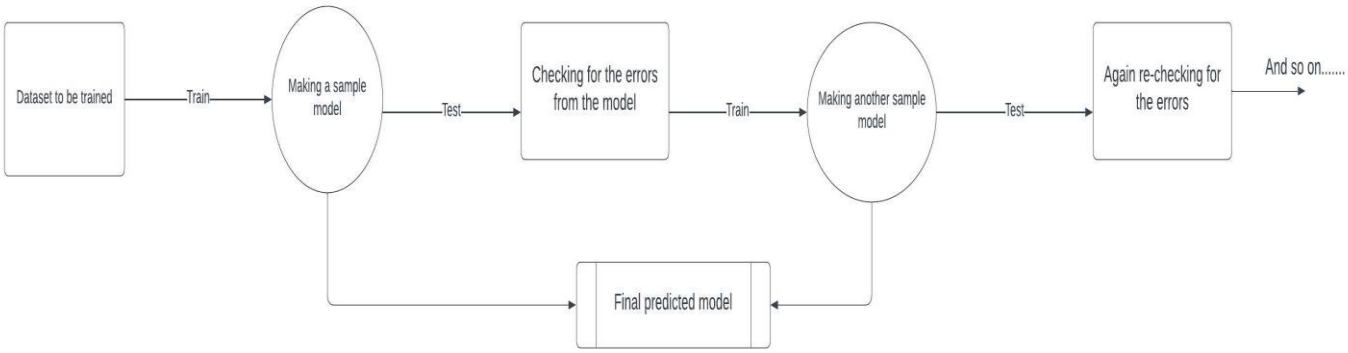


Figure 3: Gradient Boosting

4.2 Random Forest Regressor

Random Forest is a widely used machine learning algorithm for both classification and regression-related problems. It falls under the umbrella of supervised learning techniques and works on the method of ensemble learning. By combining multiple classifiers, Random Forest aims to tackle complex problems and enhance model performance.

In this project, Random Forest Regressor has been used to determine the insulin dosage. As the name implies, Random Forest Regressor comprises numerous decision trees trained on various subsets of the dataset, leveraging their collective predictions to improve overall accuracy. Unlike relying on a single decision tree, this approach aggregates predictions from each tree and determines the final output based on majority voting.

The process involves several steps:

1. Randomly selecting K data points from the training set.
2. Constructing decision trees using the chosen data points (subsets).
3. Specifying the desired number N of decision trees to build.
4. Iterating through steps 1 and 2.
5. When presented with new data points, determine the predictions of each decision tree and assign the new data points to the category that garners the most votes.

This method effectively harnesses the diversity of decision trees to improve the robustness and predictive power of the model, making Random Forest a popular choice across various machine learning applications.

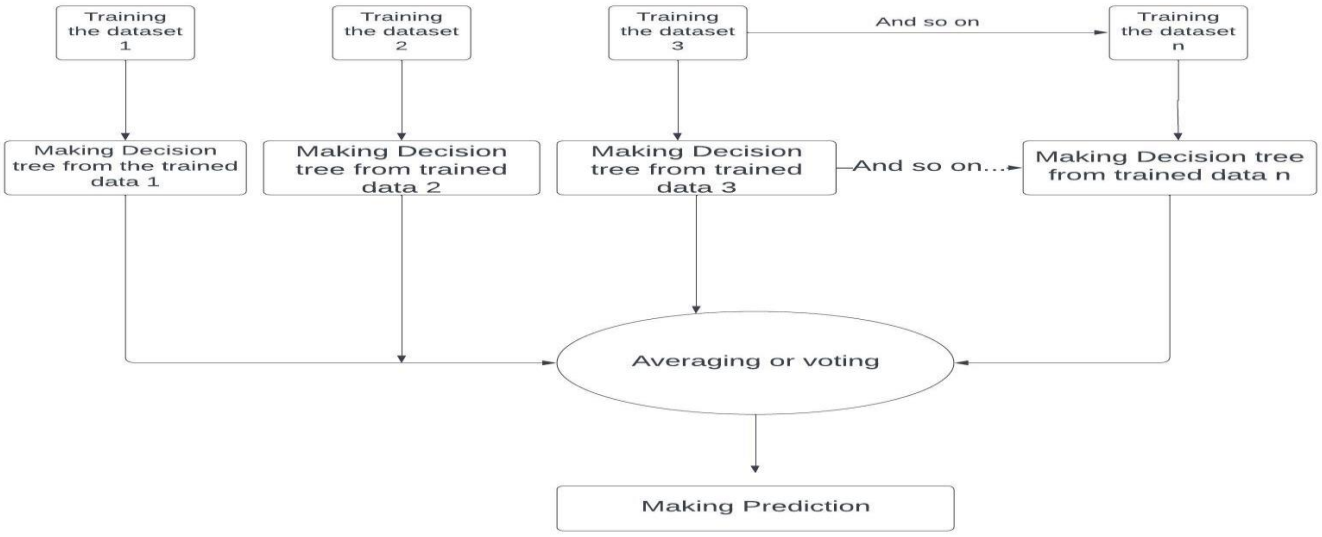


Figure 4: Random Forest Regressor

V. RESULTS AND DISCUSSION

The proposed system aims to assess the severity levels detected in diabetic patients through the application of diverse processing techniques coupled with machine learning algorithms. By using advanced computational methods, the system endeavors to categorize diabetic cases based on their severity levels. This involves analyzing patient data using machine learning algorithms to uncover patterns and relationships that contribute to the determination of severity. The outcomes of the system will provide valuable insights into the severity of diabetic conditions, facilitating better understanding and management of the disease for healthcare practitioners and patients alike.

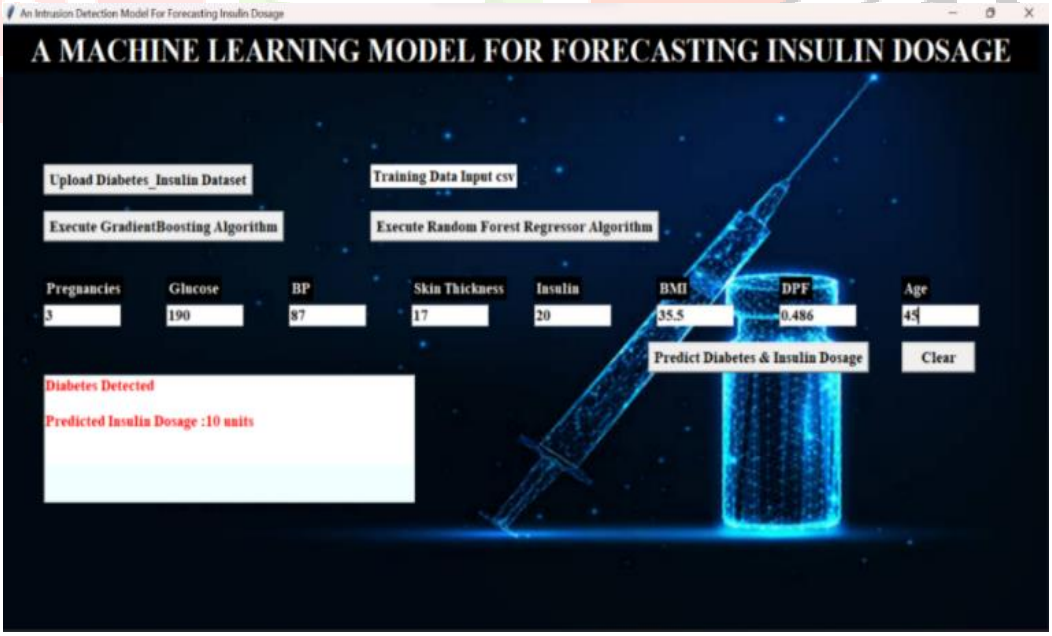


Figure 5: Output Screen

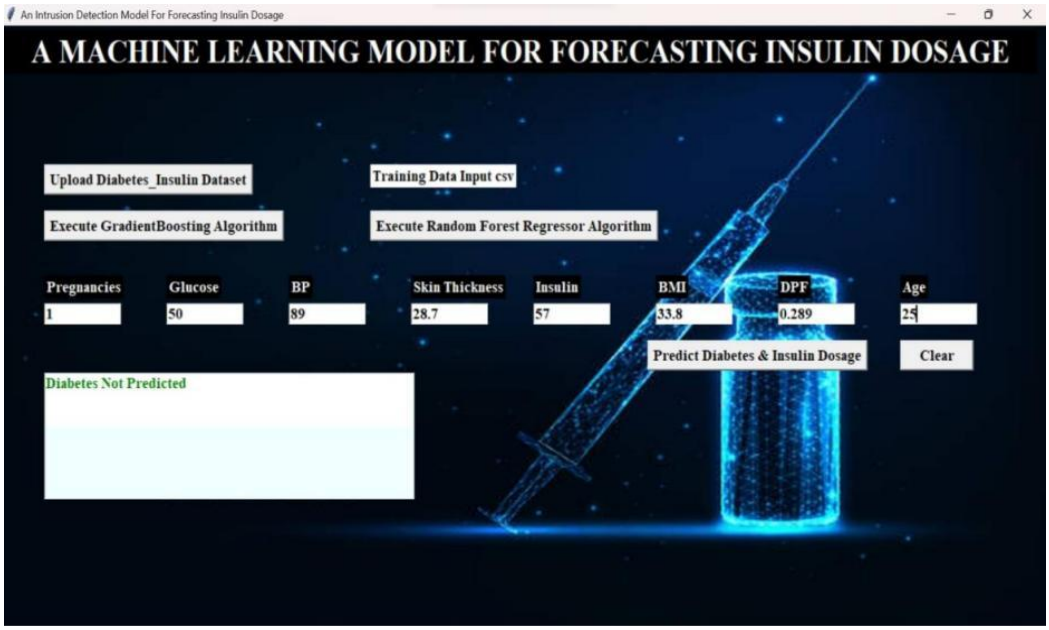


Figure 6: Output Screen

REFERENCES

[1] American Diabetes Association. "Diagnosis and Classification of Diabetes Mellitus." Diabetes Care, Vol. 31, No. 1, 2008, pp. 55-60.

[2] I. Eleni Georga, C. Vasilios Protopappas, and I. Dimitrios Fotiadi. "Glucose Prediction in Type 1 and Type 2 Diabetic Patients Using Data-Driven Techniques." Knowledge-Oriented Applications in Data Mining, 2011, pp. 277-296.

[3] V. Tresp, T. Briegel, and J. Moody. "Neural-Network Models for the Blood Glucose Metabolism of a Diabetic." IEEE Transactions on Neural Networks, Vol. 10, No. 5, 1999, pp. 1204-1213.

[4] C. M. Bishop. Pattern Recognition and Machine Learning. Springer, 2006, New York.

[5] S. Haykin. Neural Networks and Learning Machines. Pearson, 2008

[6] W. D. Patterson. Artificial Neural Networks - Theory and Applications. Prentice Hall, Singapore, 1996.

[7] M. Pradhan and R. Sahu. "Predict the Onset of Diabetes Disease Using Artificial Neural Network (ANN)." International Journal of Computer Science & Emerging Technologies, Volume 2, Issue 2, April 2011, p. 303.

[8] W. Sandham, D. Nikoletou, D. Hamilton, K. Paterson, A. Japp, and C. MacGregor. "Blood Glucose Prediction for Diabetes Therapy Using a Recurrent Artificial Neural Network." EUSIPCO, Rhodes, 1998, pp. 673-676.

[9] Gradient Boosting – AI Wiki – accessed on February 17,2024 - https://images.app.goo.gl/TNky9YA1tZswNyym7

[10] Random Forest – Javapoint – accessed on February 16,2024 - https://images.app.goo.gl/bzjLbEWoi9XhwhTR8

[11] PIMA Diabetes Dataset – Kaggle – accessed on December 14,2023 - https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

[12] UCI Insulin Dosage Dataset – UC Irvine Machine Learning Repository – accessed on January 20,2024 - https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database