International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 4 April 2023, pp: 877-883 www.ijaem.net ISSN: 2395-5252

Prediction of Diabetes and Insulin Dosage

V Anirudh, G Siri, Puli Shashank, T Rahul Vardhan, Veera Bhadra Rao

Dept of Computer Science and Engineering GITAM (Deemed to be University) Visakhapatnam, India
Dept of Computer Science and Engineering GITAM (Deemed to be University) Visakhapatnam, India
Dept of Computer Science and Engineering GITAM (Deemed to be University) Visakhapatnam, India
Dept of Computer Science and Engineering GITAM (Deemed to be University) Visakhapatnam, India
M.Tech, Assistant Professor, Dept of Computer Science and Engineering GITAM (Deemed to be University) Visakhapatnam, India

Date of Submission: 10-04-2023 Date of Acceptance: 20-04-2023

ABSTRACT— The world has different types of people as it differs with health issues. In that one of the main issues which they agonize is Diabetes. Age, obesity, hereditary habits, bulimia, elevated blood pressure, and others can contribute to Diabetes Mellitus. Diabetes boosts a person's chance of expanding various illnesses, heart disease, kidney disease, heart attack, eye issues, nerve damage, etc. Multiple trials are used in hospitals to gather the data required to diagnose Diabetes by taking appropriate treatment. Data which is extracted in the healthcare sector are massive. Big Data analyzes extensive data. It can be learned from the model and obtain accuracy for further analysis.

I. INTRODUCTION

Healthcare is very crucial in our life. Considering the current scenario, in developing countries like India, Diabetic Mellitus (DM) has become a very severe disease. Diabetic Mellitus (DM) can distinguish as Non-Communicable Disease (NCB), and many people suffer.

- About 90-95 percent of adult cases are Type-2 Diabetes. It has been circulating to all age groups without affecting severe conditions. In the present scenario, Diabetes is prevalent.

Type-2 is also known as Non-Insulin-Dependent Diabetes Mellitus (NIDDM). This type of Diabetes is seen when body cells cannot use insulin properly.

- Diabetes Mellitus (DM) is categorized as Type-1 or Insulin- Dependent Diabetes Mellitus (IDDM). The inability of the human body to generate sufficient insulin is the reason behind this type of DM; hence, it is required to inject insulin into a patient.

1.1 WHAT AGE DOES IT OCCUR?

Healthcare is very crucial in our life. Considering the current scenario, in developing countries like India, Diabetic Mellitus (DM) has become a very severe disease. Diabetic Mellitus (DM) can distinguish as Non- Communicable Disease (NCB), and manypeople suffer.

About 90-95 percent of adult cases are Type-2 Diabetes. It has been circulating to all age groups without affecting severe conditions. In the present scenario, Diabetes is prevalent. Type-2 is also known as Non-Insulin-Dependent Diabetes Mellitus (NIDDM). This type of Diabetes is seen when body cells cannot use insulin properly.
Diabetes Mellitus (DM) is categorized as Type-1 or Insulin- Dependent Diabetes Mellitus (IDDM). The inability of the human body to generate sufficient insulin is the reason behind this type of DM; hence, it is required to inject insulin into a patient.

II. RELATED WORKS

There have been several related works have been done earlier. We are creating a new method involving earlier algorithms with a new technology that deploys the method. For the Pima Indian diabetes data set, machine learning algorithms are applied in the Hadoop Map Reduce environment to identify missing values and identify patterns. This research will be able to forecast common kinds of Diabetes, risks in the future, and the types of treatments that can be given depending on the risk level of the patient.

Here, a glimpse of its applications in various fields is also addressed by emphasizing its importance in the healthcare sector. From this, we also took research about the diagnosis of Diabetes, which was done using classification mining techniques. From recent news articles or references,

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 4 April 2023, pp: 877-883 www.ijaem.net ISSN: 2395-5252

we learned that over 246 million people around the globe have Diabetes, with women making up the majority of those affected. The WHO study predicts that by 2025, this figure will have increased to over 380 million. With no immediate treatment, the illness has been ranked the fifth most lethal in the US. The prevalence of Diabetes and its symptoms has increased along

with the development of information technology and its ongoing entry into the medical and healthcare fields. This study uses the Decision Tree and Naive Bayes algorithms to classify the data to identify patterns that can be used to diagnose the illness.

The study aims to suggest a quicker and more effective method of disease diagnosis that will enable patients to receive therapy immediately.

III. METHODOLOGIES

The analysis from several machine learning methods follows. The algorithms adopted in this study are Logistic

Regression, Random Forest, Xgboost, Decision Tree Regression, and Ada Boost. The reports are valuable for experts or medical analysis to identify whether the following patients where we take patients as women.

As this total procedure approaches Diabetes and insulin dosage using various algorithms. The data was collected from multiple sources and a clinic that was compared and checked with the error. We have segregated this data into two sections named "Training" and "Testing" in the ratio of 80:20.

1.1 Xg Boost

XgBoost is a distributed grade-boosting library optimized for effective and scalable training of machine literacy models.

It is an ensemble literacy system that integrates the models that are weak learners to obtain an accurate prediction.

1.2 Logistic Regression

Categorical dependent variables that may obtain a conclusion have been announced by performing logistic regression. Hence, the result must be a discrete or absolute value. The output value gives us the probabilistic deals between 0 and 1 rather than producing an accurate value. This choice is either Yes or No, 0 or 1, True or False, etc.

P is the Probability of both input values x,y where y is a dependent variable and x is an independent variable.

P(y = 1|x) and z is the linear combination of variables x and their coefficients, and exp is the exponential function.

1.3 Ada Boost

The AdaBoost algorithm, also called adaptive boosting, is a boosting technique used in machine learning as an ensemble approach. Every time, a new set of weights is utilized, with instances that were incorrectly classified receiving more weight.

1.4 Random Forest

Random Forest is a supervised learning approach employed in the machine learning algorithm. It may be applied to solve categorization and regression challenges. The concept of ensemble learning is a procedure that combines unique classifiers to address a complicated problem and enhance the model's performance.

Y =

f(x)(x) where:

Based on the input variables x, y is the expected output variable (or class).

The decision tree function f(x) uses x as an input parameter and returns y as the expected result.

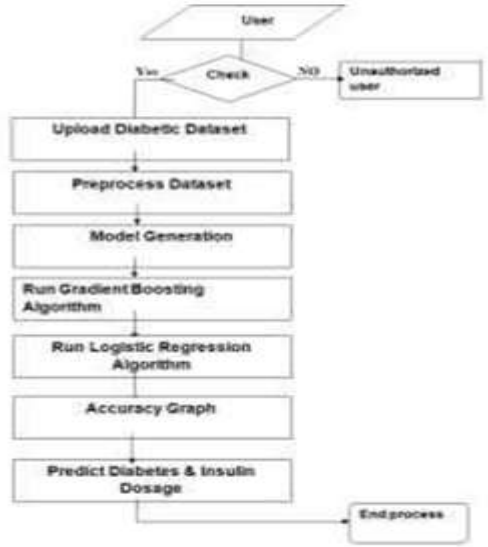
1.5 Decision Tree Regression

This machine learning data structure divides a data set into smaller subsets based on their characteristics.

The concept of repeatedly dividing the records where the data is gathered is called a decision tree. Until the data is segregated, the differentiation of the concept is limited to just one class. For instance, a tree where each branch differs from a collection of Yes or No questions, categorizing the data at each point.

IV. FLOWCHART

1.6 Flowchart of the Model



International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 4 April 2023, pp: 877-883 www.ijaem.net ISSN: 2395-5252

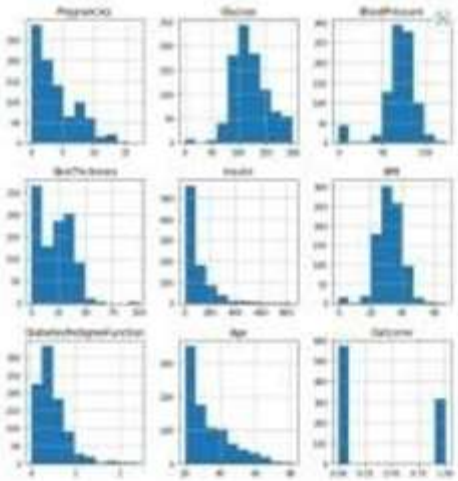
1.7 Proposed System Work Flow

4.1 System work flow

- Develop a predictive model: The primary objective of the project is to develop a predictive model that accurately predicts a patient's blood glucose levels andinsulin dosage requirements based on various patient data such as age, gender, BMI, physical activity level, diet, medication, and other relevant data.
- Improve diabetes management: The project aims to improve diabetes management by providing healthcare professionals with a more accurate and personalized approach to insulin dosage and blood glucose level prediction. This can help reduce the risk of complications and improve patient outcomes.
- Reduce healthcare costs: By improving diabetes management, the project aims to reduce healthcare costs associated with diabetes treatment. This can be achieved by reducing hospitalizations, emergency room visits, and other healthcare costs associated with diabetes- related complications.
- Increase patient adherence: The project aims to increase patient adherence to insulin therapy by providing a personalized and accurate approach to insulin dosage. This can help patients manage their Diabetes more effectively and reduce the risk of complications.
- Personalized medicine: The project aims to provide personalized medicine to diabetic patients by considering various patient data such as age, gender, BMI, physical activity level, diet, medication, and other relevant data. This can help healthcare professionals tailor treatment plans for each patient's needs.
- Early detection of Diabetes: The project aims to detect Diabetes early by analyzing patient data such as age, gender, BMI, physical activity level, diet, medication, and other relevant data. Early detection of Diabetes can help prevent complications associated with the disease.
- Improve patient education: The project aims to improve patient education by providing patients with personalized information about their diabetes management. This can help patients make informed decisions about their treatment and improve their understanding of the disease.
- Better communication between healthcare professionals and patients: The project aims to improve communication between them by providing them with accurate and personalized information about diabetes management. This can help patients better understand their treatment and improve their adherence to insulin therapy.
- Integration with Electronic Health Records (EHRs): The project aims to integrate the predictive model with EHRs to provide healthcare professionals easy access to patient data. This can help healthcare professionals make informed decisions about treatment plans and improve patient outcomes.
- Real-time monitoring: The project aims to monitor blood glucose levels and insulin dosage requirements. This can help patients and healthcare professionals adjust treatment plans to prevent complications.

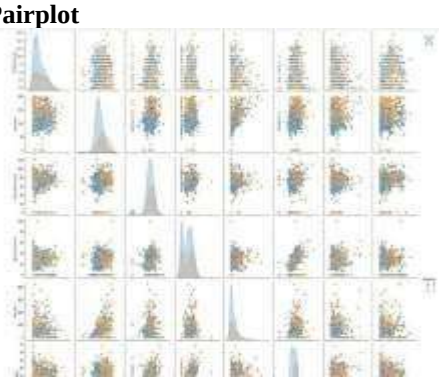
V. GRAPHICAL REPRESENTATION

1.8 Histogram of Data



5.1 Histogram of all the attributes contained in the data.

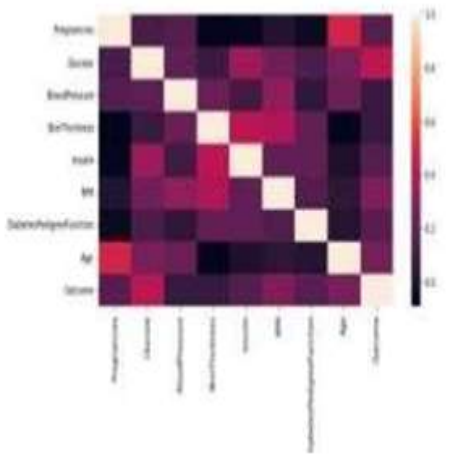
5.2 Pairplot



1.9 Pairplot of all the data which contains the outcomeof two things, i.e., 0 and 1.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 4 April 2023, pp: 877-883 www.ijaem.net ISSN: 2395-5252

1.10Heatmap



5.3 Heatmap of all the attributes showing independent and dependent values

VI. IMPLEMENTATION

- **Data Collection:** The first step in developing the prediction model for Diabetes and insulin dosage is collecting data. This can involve gathering data from various sources, including electronic health records, patient monitoring devices, and patient-reported data. The data should include relevant patient information such as age, gender, weight, height, and medical history.

- **Data Pre-processing:** After collecting the data, the next step is to pre-process it. This can involve cleaning the data by removing missing values, outliers, and irrelevant features. Data normalization or scaling is also necessary to ensure the data is comparable.

- **Feature Selection:** Feature selection selects the most relevant features from the pre-processed data. This step is crucial to reduce the complexity of the model and improve its performance. Feature selection can be made using various techniques such as correlation analysis, principal component analysis, and mutual information- based feature selection.
- **Model Development:** The next step is to develop the prediction model after selecting the features. Various machine learning techniques are available for develop- ing the model, such as decision trees, support vector machines, and artificial neural networks.

The model should be trained using the pre-processed data, and its performance should be evaluated using appropriate metrics such as accuracy, sensitivity, and specificity.

- **Model Optimization:** Once the model is developed, it may require optimization to improve its performance. This can involve tweaking the model's parameters, adjusting the feature selection technique, or changing the machine learning algorithm used to develop the model. Optimization should be iterative and data-driven to achieve the best possible performance.
- **Model Validation:** After optimizing the model, it must be validated using independent data. This step is essential to ensure the model can generalize to new data and not overfit the training data. The model's performance should be evaluated using appropriate metrics, and its accuracy and reliability should be validated.

- **Implementation:** After the model is developed, optimized, and validated, it needs to be implemented in a real-world setting. This can involve integrating the model into healthcare systems or mobile applications that patients can use to monitor their diabetes and insulin dosage. The implementation process should be carefully planned and executed, considering the ethical and legal issues associated with data privacy and security.

- **Continuous Improvement:** Finally, the model should be continuously monitored and improved. The model should be retrained and optimized to improve its performance as new data becomes available. Additionally, feedback from patients and healthcare providers should be incorporated into the model's development and optimization process to ensure that it meets their needs and expectations.

VII. TABLES AND REPRESENTATION

The table represents all the attributes from the data where contains Pregnancy, Glucose, Blood Pressure, Skin Thickness, Insulin, Body Mass Index, Diabetic Pedigree Function, Age, and Outcome. Each attribute has its variation and description where it represents any one of the attributes is higher then, it leads to dangerous diseases like Diabetes and heart attack, etc.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 4 April 2023, pp: 877-883 www.ijaem.net ISSN: 2395-5252

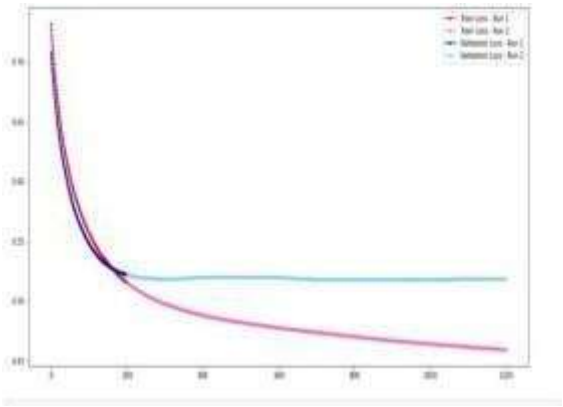


Figure 8 Graph between Validation and Train Loss

Pregnancies	Glucose	BP	Skin Thickness	Insulin	BM I	PDF	Age	Outcome	Output
7	100	80	20	35	30	0.484	30	1	Yes
0	118	84	47	230	45.8	0.551	31	1	Yes
2	197	45	60	55	40.5	1.5	28	0	No
1	189	60	23	570	55	0.39	25	1	Yes
10	115	120	0	0	26.9	1.4	34	0	No
5	117	92	54	0	34.2	0.337	38	0	No
4	103	60	33	192	24	2.9	33	0	No
5	109	75	26	0	36	0.54	60	0	No
8	99	84	0	0	35.4	0.388	50	0	No
3	138	76	36	245	31.6	0.851	48	1	Yes

VIII. EVALUATION

The process of this study or the project shows that few other studies can also be treated in other ways, such as using other models to predict the output. First, the data should be collected from the sources. To get the precise outcome we have taken the data from a clinic where patients take blood tests to a maintain a good and balanced life without no health issues.

As the diabetes is very common disease for anyone without

IX. RESULTS

The dataset contains of 887 records in which 80% has Training and 20% has Testing set. The output shows the dataset needs to be uploaded as the dataset contains two different data one is Diabetes comparing the age the factor. To show case the results we have and other is Insulin. After uploading he dataset, the screen shows done the process for both the datasets. We came to know that the data has been uploaded and next thing is to pre-process the data. these types of input values or attributes in data are very crucial so, we tried it using neural network as building a single hidden layer model. We compared both the datasets and found out the error loss. As the error loss must be less and we got 0.2-0.5%.

After pre-processing, we need to perform the following algorithms, i.e., Gradient Boosting and Logistic Regression. Both the algorithms performed will show the accuracy in the model. Next, the accuracy graph as the pair plot shows all the attributes of the data which We performed data cleaning where no NaN values present though contains the patients having diabetes or not. It is written in we have taken two data, i.e., Diabetes

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 4 April 2023, pp: 877-883 www.ijaem.net ISSN: 2395-5252

and Insulin Dosage. We categorical values 0 and 1 with two different colours orange and have tried several models to build, but the algorithms performed blue. 0 is not detected and 1 is detected. The final step to finish this are not satisfying as the data needs to be cleaned. By plotting the study is to predict the values. As everyone takes the recent values of graph we can consider what corrections need to be done. For each the data and concludes with the output but, we take the test values to attribute we plotted histogram and heatmap to check whether it get whether the input of that data consists is right or wrong. The test conforms of any missing or independent values are taken. In values are taken and predicts the values and concludes the output neural networks the main thing arises is about the error values and with few values and tells that if the data is sufficient then there is no accuracy. By giving higher epochs with different numbers it may detection of diabetes. If it not satisfied then it shows diabetes be successful. Epoch values 200 and 1000 may get a lot of detected and tells the value of insulin dosage to use.difference in this model. The testing starts with two different sections, i.e., "Training" and "Testing". We have decided to contribute 80% in training and 20% in testing as we proceed forward with logistic regression shows to 72 – 75%. After using Hyper parameter Optimization where the model gets trained and tries to improve the older model and get the better results. 80% of Accuracy has been shown and tried using Gradient Boosting, it showed to 98% of accuracy. There it comes the graph about which include validation and train loss. Two colors has been chosen for each of the loss which represents how the network of our model is optimized or it has enough training as we increased the epoch size. The graph shown below represents both the losses, i.e., Validation Loss and Train Loss .As we can see the Training Loss is varying differently compared to Validation Loss, Training Loss goes deeper where it looks like the Validation Loss has stabilized (or even gotten worse!). This suggests that our network will not benefit from further training.

1.11 ROC CURVE

9.1 represents ROC Curve for output value

X. CONCLUSION

This project's primary objective was to use the Gradient Boosting Classifier and Logistic Regression algorithms to predict diabetes and insulin dosage in diabetic-detected patients. For this, we need to accomplish the goal, i.e., PIMA diabetes and UCI Insulin dosage datasets were chosen to train both algorithms.

Subsequently, the test dataset, without any class label, was The project outcomes demonstrate that both algorithms can be practical tools for diabetes prediction and insulin dosage. However, future research could focus on improving the accuracy of the algorithms and expanding the dataset to improve the generalizabilityof the results.

XI. FUTURE SCOPE

To further improve the accuracy of our model for predicting diabetes and insulin dosage, we plan to integrate additional methods for parameter tuning. By incorporating these techniques, we aim to enhance the precision of our model and provide more reliable predictions. To achieve this, we will explore various methods for parameter tuning, such as grid search and Bayesian optimization.

These approaches will enable us to search for our model's optimal set of parameters, improving its performance and accuracy. To ensure the reliability of our model, we will also conduct extensive validation tests using cross-validation and other techniques. This will enable us to identify potential overfitting or underfitting issues and

ensure that our model generalizes well to new data. Furthermore, we plan to explore machine-learning algorithms and ensembles to determine which approach provides the best results for predicting diabetes and insulin dosage. By comparing and contrasting different techniques, we can gain more insights into how our model works and identify potential areas for improvement.

Finally, we will conduct a thorough analysis of the performance of our model and compare it with other state-of-the-art models in the uploaded. The Gradient Boosting algorithm predicted the presence literature. This will enable us to evaluate the effectiveness of our of diabetes, and the Logistic Regression algorithm predicted insulin dosage only if Gradient Boosting detected diabetes. The model and determine whether it provides better accuracy than existing approaches. In summary, our future work will focus on project outcomes exhibited that both algorithms performed well in enhancing the accuracy of our model for predicting diabetes and predicting diabetes and insulin dosage. The Gradient Boosting algorithm achieved an accuracy of 95-98% in predicting the presence of diabetes, while Logistic Regression attained an accuracy of 75-80% in predicting insulin dosage. These results suggest that the two algorithms can effectively predict diabetes insulin dosage by integrating additional methods for parameter tuning, testing our model using a larger dataset, incorporating additional features, and exploring different machine learning algorithms and ensembles. By conducting extensive validation tests and comparing our model's performance with other approaches, we and insulin dosage in diabetic-detected patients. The project hope to provide better recommendations for patients with diabetes trained the Gradient Boosting Classifier and Logistic Regression algorithms using these two datasets. The Gradient Boosting algorithm employs an ensemble of weak learners to improve the overall model's accuracy. The Logistic Regression algorithm is a widely used classification algorithm that uses a logistic function to predict the probabilities of binary outcomes. The project results demonstrate that both algorithms can accurately predict diabetes and improve their health outcomes.

REFERENCES

- A. Belle, R. Thiagarajan, S. M. R. Soroushmehr, F. Navidi, D. A. Beard, and K. Najarian, “Big Data Analytics in Healthcare,” Hindawi Publ. Corp., vol. 2015, pp. 1–16, 2015.
- J. Andreu-Perez, C. C. Y. Poon, R. D. Merrifield, S. T. C. Wong, and G.-Z. Yang, “Big Data for Health,” IEEE J. Biomed. Heal. Informatics, vol. 19, no. 4, pp. 1193–1208, 2015
- E. Ahmed et al., “The role of big data analytics in Internet of Things,” Comput. Networks, vol. 129, no. December, pp. 459–471, 2017
- “The big-data revolution in US health care: Accelerating value and innovation | McKinsey & Company.” [Online]. Available: https://www.mckinsey.com/industries/healthcare-systems-andservices/our-insights/the-big-data-revolution-in-us-health-care. [Accessed: 12-May-2018].
- M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, “Disease Prediction by Machine Learning over Big Data from Healthcare Communities,” IEEE Access, vol. 5, no. c, pp. 8869–8879, 2017.
- L. Zhou, S. Pan, J. Wang, and A. V. Vasilakos, “Machine learning on big data: Opportunities and challenges,” Neurocomputing, vol. 237, pp. 350–361, May 2017.