JOURNAL OF CRITICAL REVIEWS

ISSN- 2394-5125 VOL 7, ISSUE 19, 2020

Dia-Analyze: A Holistic Suite for Data Analytics in Type 2 Diabetes for Advanced medical applications

K. Sivanagireddy
Professor, Department of Electronics and Communication Engineering,
Sridevi Women's Engineering College, Hyderabad, India,
sivanagireddykalli@gmail.com

Abstract- Tailoring long-term care for individuals with chronic conditions such as Type 2 Diabetes (T2D) is imperative due to the variability in responses observed among patients, even when undergoing identical treatments. Analyzing extensive patient data, often referred to as "big data," presents a promising avenue for studying the diverse manifestations and impact of T2D, utilizing the wealth of digitized patient records. The field of data science can significantly contribute to the customization of care plans, validation of established medical knowledge, and discovery of valuable insights within the extensive healthcare datasets. This comprehensive review introduces a framework for effectively managing T2D, covering various stages, including exploratory analysis, predictive modeling, and visual data exploration techniques. This integrated approach empowers healthcare professionals and researchers to identify meaningful correlations between a patient's diverse biological markers and the complications associated with T2D. By utilizing this framework, it becomes possible to predict how an individual will respond to specific treatments, categorize T2D patients into distinct profiles associated with particular conditions, and assess the likelihood of complications linked to T2D.

Keywords – Type 2 Diabetes (T2D). Machine Learning and AI with medical applications.

1. INTRODUCTION

In recent times, various industries, including healthcare, have witnessed a significant rise in the pursuit of data-driven solutions. This enthusiasm can be attributed to the swift progress in cloud technologies, substantial data frameworks, and artificial intelligence. Nevertheless, the establishment of expansive data systems, like applications for healthcare data analytics, demands a careful approach involving precise design, thoughtful planning, and a strong partnershipbetween healthcare experts and pertinent stakeholders.

This is crucial due to the sensitive nature of healthcare data and its potential impact on patient well-being. To address this, the EU assigned AEGLE with the task of developing a robust big data system aimed at providing extensive data services to the healthcare industry.

These services encompass data analysis, storage of electronic health records, utilization of cloud services to accelerate processing for complex analytics, and real-time handling of large volumes of data. A detailed depiction of the AEGLE environment can be found in Figure 1.

The AEGLE initiative has formulated an all-encompassing strategy detailed in [1]. Within the framework of the AEGLE project, numerous data studies, including investigations into Type 2 Diabetes (T2D), have been conducted. T2D stands as an increasingly prevalent chronic ailment, serving as a widespread contributor to health complications and mortality, while also exerting substantial pressure on healthcare resources. According to Public Health England (PHE) records from 2015, T2D impacted 3.8 million adults aged 16 and above in England,a figure that was anticipated to escalate to 4.7 million by 2019. The world health Organization (WHO) ranks T2D as the seventh principle cause of death on a global scale.

In United States, diabetes is approximated to generate expenses amounting to $327 billion, thereby yielding a significant economic consequence [5]. As a result, it becomes crucial to implement efficacious treatment methods and initiate timely interventions to alleviate the influence of T2D on patients' well-being and financial burdens. Starting from the 1980s, there has been a notable upsurge in the digital documentation of patient information. This extensive collection of healthcare records presently empowers data specialists to scrutinize and unveil previously undiscovered trends and connections, potentially advancing our comprehension of illnesses and their management.

ISSN- 2394-5125 VOL 7, ISSUE 19, 2020

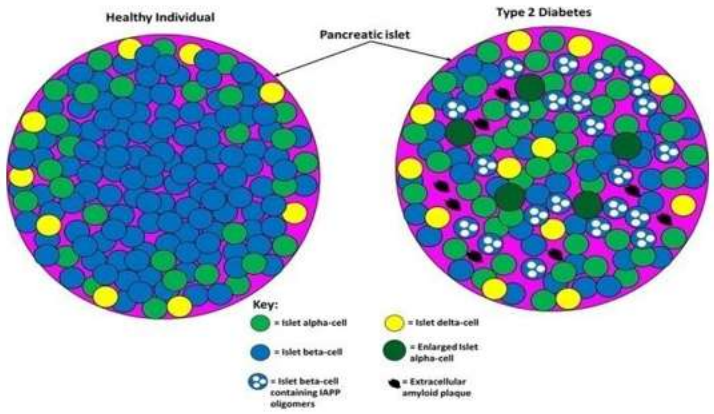

Figure 1: model

The above Figure is extracted from the base paper which demonstrates the differences between Type -2 -diabetes, & Non-diabetes.

By harnessing historical data derived from a cohort of patients, scientists can construct models that prognosticate the trajectory of a patient's ailment and adapt their treatment regimen accordingly [6]. A multitude of research endeavors have concentrated on the realm of data analysis pertaining to Type 2 Diabetes (T2D). Notably, a particular facet of T2D that has garnered attention is the forecasting of complications. Diverse models have been employed for this purpose, spanning from classical Cox's models and their iterations [7-9] to more contemporary machine learning-based techniques such as support_vector_machines (SVM) [11], Bayesian methodologies [12], nearest neighbor approaches [13], random_forest_algorithms [14],

logistic_regression_models [15], genetic-algorithms [16], and deep_learning_ methodologies[17-19]. The broad spectrum of models formulated via thorough analysis of T2D data possesses the capacity to aid healthcare practitioners in comprehending data and making informed choices. This article outlines our endeavors in scrutinizing T2D data with the objective of predicting patient responses to treatments, uncovering associations among distinct patient attributes, and evaluating the likelihood of diverse complications. This work signifies an initial stride toward establishing a unified T2D analysis toolkit, engineered to educate students and professionals regarding theailment and its management.

2. LITERATURE REVIEW

DIMITRIOUS SOUDRIS [1] The AEGLE project has set forth the objective of creating an innovative information technology solution spanning the entirety of the healthcare data value chain. This solution will be constructed by harnessing cloud computing technologies, which encompass high-performance computing (HPC) platforms, dynamic resource-sharing mechanisms, and cutting-edge visualization approaches. This article delves into the domains of Big Data healthcare settings that have been tackled, in addition to highlighting the pivotal enabling technologies. Furthermore, the discussion extends to encompass considerations related to information security and regulatory aspects that are integral within the AEGLE framework. The assimilation of such technological strides stands to yield notable advantages in the realm of advanced healthcare analysis and its interconnected research endeavors.

J.M.M RUMBOLD [1] Reflect on the current and future possibilities that Big Data offers in the realm of diabetes management. Undertake a comprehensive review of scholarly literature focusing on the intersection of diabetes care and Big Data. The outcomes of this exploration underscore the transformative potential of the rapidlygrowing healthcare data landscape in reshaping diabetes care. Notably, the influence of Big Data is already beginning to shape diabetes treatment through meticulous data analysis. Nevertheless, conventional healthcare methodologies

have yet to unlock the complete potential of Big Data. A phase will emerge when this integration becomes commonplace. Acknowledging the substantial volume of healthcare data being amassed and the consequential value of extracting insights for improved care is essential. However, it is crucial to acknowledge that substantial developmental efforts areessential to realizing these aspirations.

CAROL COUPLAND [2] The central research inquiry revolves around the feasibility of formulating algorithms capable of predicting the susceptibility to visual impairment and lower limb amputation in individuals aged 25 to 84 who have diabetes, spanning a span of 10 years.

The investigation utilized data from approximately managed healthcare facilities in England spanning the years 1998 to 2014. These data inputs were drawn from the Q Research and (CPRD) databases. The construction and validation of the models were conducted using data from 254 Q Research practices (comprising 142,419 diabetes patients) and 357 CPRD practices (encompassing 206,050 diabetes patients). Moreover, an additional dataset from 763 Q Research practices (with 454,575 diabetes patients) was used for external validation purposes.

To decipher the potential for blindness and amputation risk in the next decade, Cox proportional hazards models were harnessed. These models provided diverse risk estimates for the anticipated occurrences of these complications. Calibration and discrimination metrics were employed to assess model performance across both study cohorts. The findings highlighted the development and assessment of predictive models to ascertain the absolute risk of experiencing blindness and amputation in individuals with diabetes. In the Q Research cohort, during the follow-up period, there were recorded instances of 4,822 lower limb amputations and 8,063 cases of blindness.

Consistency in risk factors was demonstrated across both study cohorts. For the external CPRD cohort, the discrimination metrics forboth amputation (D_statistic 1.69, Harrell's_C_ statistic 0.77) and visual impairment (D statistic 1.40, Harrell's C statistic 0.73) showcased strong performance. Similar results were replicated for women within the Q Research validation cohort. These algorithms bear the potential to aid healthcare practitioners in identifying patients who exhibit elevated risk levels and consequently require heightened attention or interventions.

It is crucial to underscore that these findings are predicated on available data and thus encompass inherent limitations, including the potential for incomplete data entries. Nevertheless, this study bestows valuable insights for individuals grappling with type_1 or type_2 diabetes, allowing for a more precise estimation of their likelihood of encountering these complications over the ensuing decade. Notably, the models take into account their distinctive risk profiles.

In the study by JOHN S YUDKIN [4], the focus was on addressing thelimitations of the existing Risk Equations for Complications of Type 2 Diabetes (RECODE). The objective was to develop improved equations for predicting complications. The basis for this endeavor was the dataset obtained from the Action to Control Cardiovascular Risk in Diabetes (ACCORD) study, encompassing data from 9,635 participants during the years 2001 to 2009. Additional data were drawnfrom the Diabetes Prevention Program Outcomes Study (DPPOS) with 1,018 participants from 1996 to 2001, and the Look AHEAD (Action for Health in Diabetes) study contributed data on cardiovascular and microvascular events involving 4,760 participants spanning the years 2001 to 2012. The microvascular impacts studied included neuropathy, nephropathy, and visual impairment, while the assessed outcomes encompassed myocardial infarction, stroke, severe cardiovascular failure, cardiovascular-related mortality, and all-cause mortality.

To identify predictive factors, such as demographic characteristics, clinical parameters, diseases, medications, and biomarkers, a machine learning technique known as cross-validation was employed. The newly developed risk equations were then compared to earlier models by evaluating their discrimination, calibration, and net reclassificationscore.

The study outcomes indicated strong internal and external calibration,with a slope of estimated versus observed risk ranging from 0.71 to
0.31. Additionally, moderate internal and external discrimination wasobserved, with C-statistics ranging from 0.55 to 0.84 internally and
0.55 to 0.79 externally across all scenarios.

When compared to other existing models like the UK Prospective Diabetes Study Risk Engine 2 and the

American College of Cardiology/American Heart Association Pooled Cohort Equations, the newly developed equations exhibited superior performance in identifying both microvascular and cardiovascular outcomes, as evidenced by C-statistics of 0.61 to 0.66 and slopes of 0.30 to 0.39 forfatal or non-fatal myocardial infarction or stroke.

Unlike the RECODE equations, the recently formulated risk equations offer individuals diagnosed with type 2 diabetes a more precise means of assessing their potential for complications. Financial support for this research initiative was granted by the National Institute on Minority Health and Health Disparities, the National Institutes of Health, the US Department of Veterans Affairs, and the National Institute for Diabetes and Digestive and Kidney Diseases.

Conducted by JOHN F STEINER[5], this research endeavor sought to develop and assess a predictive model concerning the six-month likelihood of severe hypoglycemic events among individuals with diabetes undergoing medication.

The development group comprised 31,674 diabetes patients who wereunder medication care at Kaiser Permanente Colorado between 2007 and 2015. In addition to this, the validation groups encompassed 12,035 HealthPartners members and 38,764 Kaiser Permanente Northwest members. The factors under consideration for inclusion within the model were sourced from electronic health records. Employing a Cox regression model capable of accommodating numerous six-month observation periods per individual, two variations of the model were created – one with 16 factors and the otherwith 6 factors.The cumulative results depicted a combined total of 850,992 six-month target periods encompassing these three cohorts. Within this span, 10,448 of these target periods witnessed the occurrence of at least one episode of severe hypoglycemia.

The model pinpointed six determinants for consideration: age, type of diabetes, HgbA1c levels, estimated glomerular filtration rate (eGFR), prior history of hypoglycemia within the preceding year, and utilization of insulin. Both prediction models displayed commendable performance. The six-variable model achieved a C-statistic of 0.81, while the 16-variable model showcased robust calibration and an impressive C-statistic of 0.84.The C-statistics observed within the external validation groups spanned from 0.80 to 0.84. To conclude, our efforts yielded the successful creation and evaluation of two distinct models designed to forecast the probability of hypoglycemia occurrence within the ensuing six months. While the simpler model may find preference under specific circumstances, it's noteworthy that the 16-variable model exhibited a slightly enhanced discrimination performance when juxtaposed with the 6-variable model.

3. METHODOLOGY

Since the 1980s, there has been a substantial increase in the electronicrecording of patient data. This vast amount of healthcare records now enables data experts to analyze and uncover previously unknown patterns and associations. Such analysis can greatly enhance our understanding of diseases and their treatment.

Researchers have developed predictive models using historical data from groups of patients, enabling them to forecast the progression of apatient's illness and design treatment plans accordingly. Various research studies have been conducted in the field of data analysis for Type 2 Diabetes (T2D).One particular area of focus in T2D research has been predicting the likelihood of complications. From the initial development of Cox's models to more recent machine learning-based models to name a few, SVM, Naïve_Bayes, nearest neighbor, Random_forest, logistic_regression, genetic algorithms, and deep learning, a variety of diverse models have been explored.

Disadvantages of the existing system:

1. The measurements introduced earlier lack innovation and holdrestricted clinical relevance.

2. Accurately forecasting disease progression and the effectiveness oftreatment interventions poses a considerable challenge.

With the progress in T2D data analysis, a requirement arises for a tool aiding healthcare experts in both data analysis and decision-making. This study aims to tackle this requirement by delving into T2D data to

anticipate patient reactions to medications, uncover associations amidst diverse patient indicators, and evaluate the potential for different complications.

This undertaking marks an initial stride towards shaping an inclusive T2D analysis toolkit, aimed at imparting knowledge to students and practitioners about the intricacies of T2D and its treatment methodologies.

Advantages of the proposed system

1. The sophisticated data analysis methodologies explored within this manuscript hold the promise of supporting physicians in making well-informed choices to elevate T2D management.

2. The metrics showcased in this article transcend limitations tied to their novelty and clinical relevance.

MODULES:

To conclude the previously discussed modules, we have organized thefollowing sections:

- Data Exploration: This tool enables us to enrich the dataset with additional information.

- Data Handling: This lesson will provide a more detailed understanding of data handling techniques.

- Data will be split into training and testing sets using this tool.

We will utilize Logistic Regression, Gaussian NB, Decision Tree, Random Forrest, ADA Boost, Gradient Boost, XG Boost

- Prediction Input: This tool will generate input for making predictions.

- At the end, the predicted number will be displayed

- Model Creation: We will utilize SVM, RF, DT, Naive Bayes, KNN,and a Voting Classifier to build the models.

- Prediction Input: This tool will generate input for making predictions.

- At the end, the predicted number will be displayed.

4. OVERVIEW OF THE DATASET

The BRFSS2015.csv dataset encompasses 70,692 survey responses
to the CDC's BRFSS2015. It maintains a balanced distribution with a50-50 split between respondents devoid of diabetes and those with
either prediabetes or diabetes.The target variable, Diabetes_binary,classifies into two categories: 0 signifies the absence of diabetes, whereas 1 indicates the presence of prediabetes or diabetes. This dataset encompasses 21 feature variables and retains a balanced
structure. The above Figure demonstrates the System Architecture.

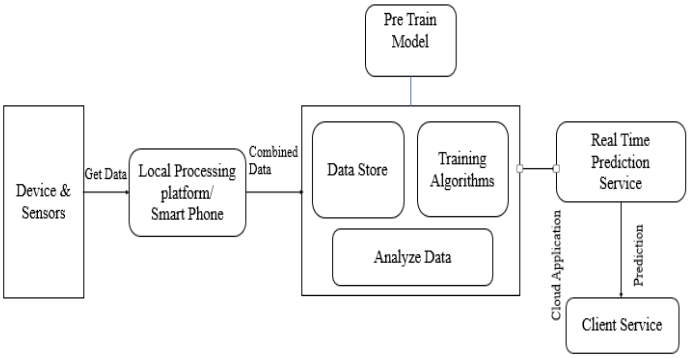

Fig.2: System architecture

5. IMPLEMENTATION

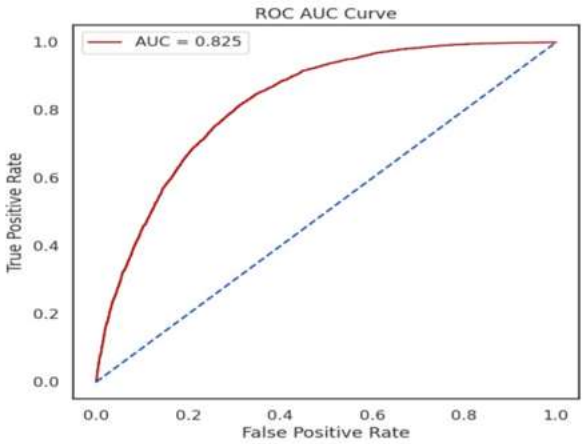

Figure 3 *LOGISTIC_REGRESSION_T2D:*

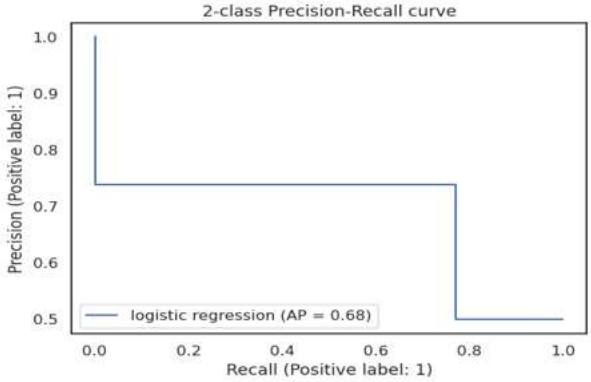

Figure 4 Recall

Logistic regression_LR: stands as a fundamental and extensively employed machine learning model designed for tasks involving binary classification. It belongs to the realm of generalized linear models and functions by forecasting the likelihood of an event's occurrence relyingon input features.

In the process of training, the model's parameters—more precisely, thecoefficients linked to the input features—are adjusted via optimizationtechniques. The goal is to reduce the dissimilarity between the projected probabilities and the factual binary labels found in the training dataset. This optimization is frequently executed using algorithms such as maximum likelihood estimation or gradient descent.

6. GAUSSIAN_NB_T2D:

Gaussian Naive Bayes (Gaussian NB) is a popular and simple machine learning model based on the Naive Bayes algorithm. It is commonly used for classification tasks, especially when dealing with continuousfeatures.

Throughout the training process, the model computes the average and standard deviation for every feature within each class. This involves determining the mean and standard deviation of individual features based on the data

ISSN- 2394-5125 VOL 7, ISSUE 19, 2020

points attributed to each respective class

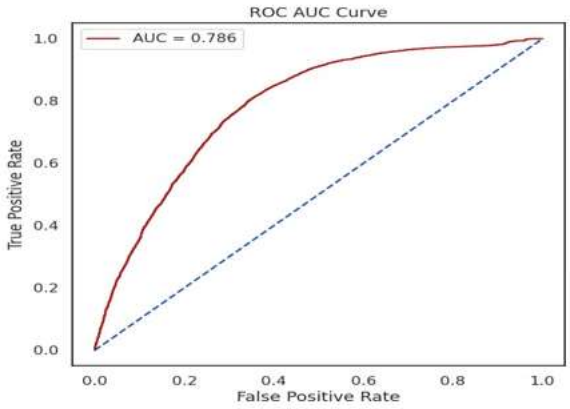

Figure 5 Roc curve

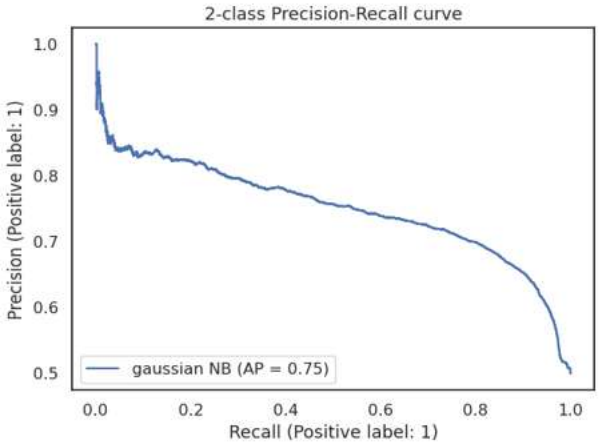

Figure 6 Recall Class

7. DECISION_TREE_T2D:

A decision tree stands as a well-recognized and easily understandable machine learning model utilized for tasks encompassing both classification and regression.

It adopts a structure akin to a tree, where each internal node signifies a choice grounded on one of the input features. In parallel, every branch corresponds to an outcome stemming from that decision, while each terminal node, or leaf node, signifies the ultimate prediction or decision.

In the realm of classification tasks, the evaluation of a node's purity is frequently accomplished using metrics like Gini impurity or entropy. Conversely, in the context of regression tasks, metrics such as mean squared error or mean absolute error are employed as measures of impurity.

Every decision tree undergoes training using a distinct random subset drawn from the training data, a method known as "bootstrapping" or "bagging." This practice entails that each tree receives training using a unique portion of the dataset, thereby introducing variability across the individual trees

ISSN- 2394-5125 VOL 7, ISSUE 19, 2020

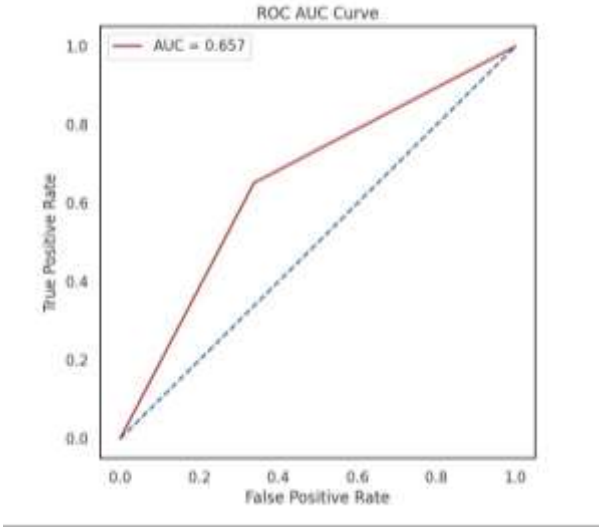

Figure: 7 ROC AUC

8. RANDOM_FOREST_T2D:

Random Forest stands as an ensemble learning technique employed in machine learning for tasks spanning classification and regression. Themethod revolves around the concept of generating numerous decisiontrees during the training phase and amalgamating their predictions to yield enhanced accuracy and resilience in predictions for fresh data.

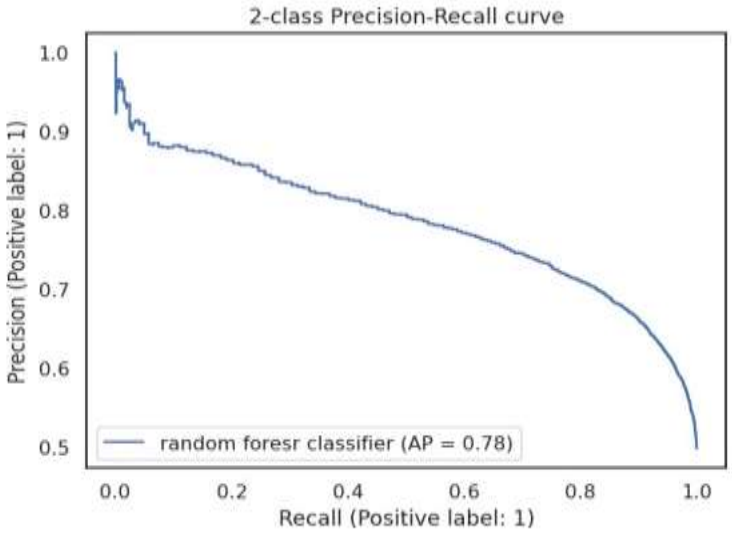

Figure: 8 ADA_BOOST_T2D:

AdaBoost, short for Adaptive Boosting, constitutes an ensemble learning technique predominantly utilized for binary classification tasks in the realm of machine learning. It is purposefully devised to elevate the efficacy of weak learners, which often encompass uncomplicated models with accuracy slightly exceeding random guesses.

This improvement is accomplished by aggregating their predictions ina weighted manner, yielding a more potent

ISSN- 2394-5125 VOL 7, ISSUE 19, 2020

and precise model. The AdaBoost algorithm operates in a series of iterations, where it sequentially trains a sequence of weak learners. During each iteration, the algorithm ascribes higher weights to incorrectly classified data points from the prior round, allowing the subsequent weak learner to place greater emphasis on the previously mishandled instances model to enhance performance in regions where its forerunner exhibited shortcomings.

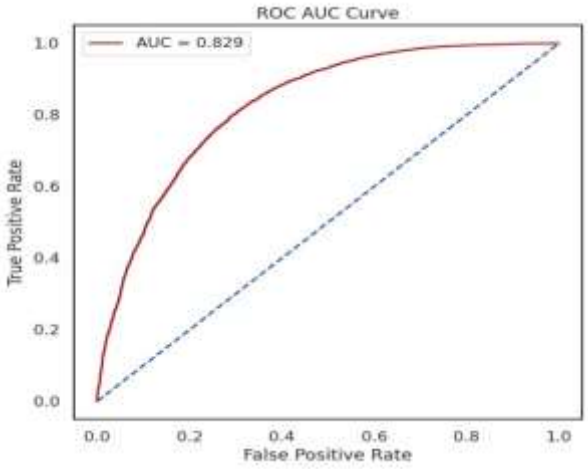

Figure 9 ROC AUC

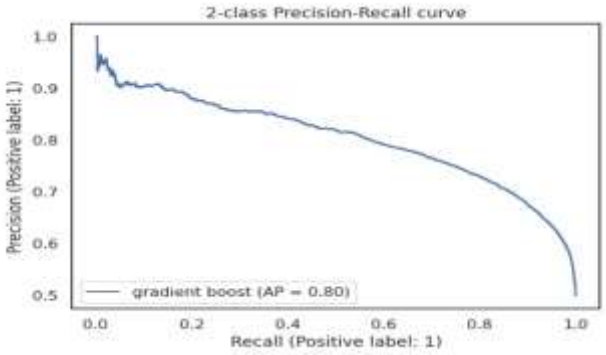

Figure : 10 GRADIENT_BOOST_T2D:

Gradient Boosting stands as an ensemble technique within the domain of machine learning, serving for both regression and classificationtasks. The fundamental premise involves amalgamating several weak XGBoost, an abbreviation for Extreme Gradient Boosting,

Unquestionably emerges as a robust and extensively adopted machine learning model categorized within the domain of gradient boosting algorithms. Initially brought to the forefront by Tianqi Chen in 2016, it rapidly garnered attention due to its remarkable efficacy and scalability. Notably, XGBoost exhibits a remarkable aptitude for managing structured/tabular data, although its utility extends to other data types like images and text as well. In machine learning competitions, XGBoost has consistently been the model of choice for many winning solutions, as it often provides state-of-the-art results. Additionally, its scalability allows it to be applied to real-world applications, such as fraud detection, customer churn prediction, recommendation systems, and more. Due to its widespread adoption and continuous development, XGBoost remains a crucial tool in the machine learning practitioner's toolkit.

ISSN- 2394-5125 VOL 7, ISSUE 19, 2020

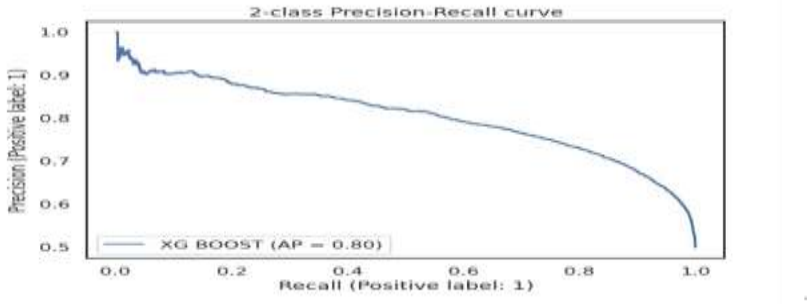

Figure : 11 2 classes precision

9. EXPERMIENTAL_RESULTS_T2D

ALGORITHM	ACCURACY	AUC	PRECISION	RECALL	F1 SCORE	MACRO AVG
LOGISTIC REGRESSION	0.74	0.82	0.76	0.73	0.74	0.75
GAUSSIAN_NB	0.71	0.78	0.72	0.72	0.72	0.72
DECISION TREE	0.65	0.65	0.66	0.65	0.65	0.66
RANDOM FORREST	0.74	0.81	0.72	0.78	0.75	0.74
ADA BOOST	0.75	0.82	0.74	0.78	0.76	0.75
GRADIENT BOOST	0.75	0.82	0.73	0.80	0.76	0.75
XG BOOST	0.75	0.82	0.78	0.80	0.76	0.75

Figure 17

The above Figure is the UI (User interphase) to predict diabetes.

ISSN- 2394-5125 VOL 7, ISSUE 19, 2020

Figure 18

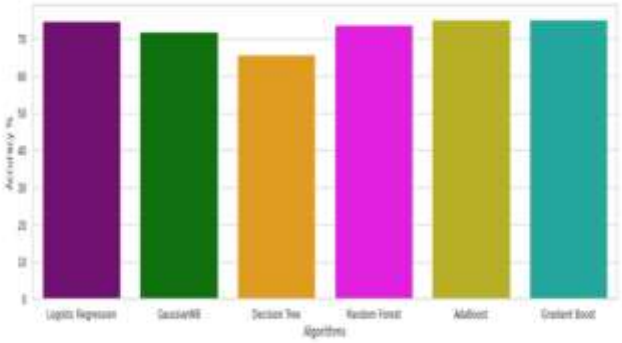

Figure 19 Prediction result.

10. CONCLUSION

Following in-depth data analysis, I proceeded to examine diverse classification models with the aim of gauging their efficacy on the dataset. The evaluation encompassed metrics such as accuracy, ROC, precision, and recall scores, yielding outcomes that met expectations. Addressing the challenge of imbalanced classification data, I implemented the SMOTE oversampling technique. My efforts didn't conclude there. I proceeded to enhance the models by conducting a Grid Search to fine-tune the hyperparameters. Subsequently, I delved into the classification report, encompassing ROC -AUC and Precision-Recall curves for each model. Upon thorough scrutiny, it emerged that Random Forest along with an array of boosting algorithms (AdaBoost, Gradient Boost, XG Boost) exhibited the most favorable alignment with our dataset. After fine-tuning the hyperparameters, the Gradient Boost Algorithm emerged as the top performer, achieving an impressive accuracy of 81.76% and an AUC of 0.834. This makes it the most suitable model for our specific task.

In this study, we presented a range of devices to explore, predict, and visualize data related to Type 2 Diabetes (T2D). We outlined three different analysis workflows: 1) Categorizing T2D patients into primary classes and identifying associations with their medical condition; 2) Constructing a predictive model to assess a patient's risk of T2D-related complications by analyzing a T2D dataset; and 3) Anticipating a patient's response to a specific treatment regimen. The results were presented more understandably, benefiting both patients and healthcare professionals due to the use of visual data representation. This empowered clinicians to make well-informed decisions about the best treatment options for T2D patients. This not only improves patient outcomes but also

ensures their safety by reducing potential side effects and speeding up recovery. The approach taken in this study represents a significant advancement in T2D management, offering a detailed and effective way to address the condition. Moreover, it has the potential to greatly benefit the healthcare system by enhancing treatment decisions and patient care. In future work, we plan to expand the dataset and train the model on larger databases to improve prediction accuracy. Additionally, we aim to develop more reliable prediction models by incorporating electronic interpretation techniques and clinically validate the findings of this study.

References

[1] D. Soudris, S. Xydis, C. Baloukas, A. Hadzidimitriou, I. Chouvarda, K. Stamatopoulos, N. Maglaveras, J. Chang, A. Raptopoulos, D. Manset, and B. Pierscionek, ''AEGLE: A big bio-data analytics framework for integrated health-care services,'' in Proc. Int. Conf. Embedded Comput. Syst., Archit., Modeling, Simulation (SAMOS), Jul. 2015, pp. 246–253.

[2] N. Holman, B. Young, and R. Gadsby, ''Current prevalence of type 1 and type 2 diabetes in adults and children in the U.K.,'' Diabetic Med., vol. 32, no. 9, pp. 1119–1120, Sep. 2015.

[3] C. D. Mathers and D. Loncar, ''Projections of global mortality and burden of disease from 2002 to 2030,'' PLoS Med., vol. 3, no. 11, p. e442, Nov. 2006.

[4] American Diabetes Association, ''Economic costs of diabetes in the U.S. in 2017,'' Diabetes Care, vol. 41, no. 5, pp. 917–928, 2018, doi: 10.2337/dci18-0007.

[5] J. M. M. Rumbold, M. O'Kane, N. Philip, and B. K. Pierscionek, ''Big data and diabetes: The applications of big data for diabetes carenow and in the future,'' Diabetic Med., vol. 37, no. 2, pp. 187–193, Feb. 2020.

[6] J. Hippisley-Cox and C. Coupland, ''Development and validation of risk prediction equations to estimate future risk of blindness and lower limb amputation in patients with diabetes: A cohort study,'' BMJ, vol. 351, no. 1, Nov. 2015, Art. no. h5441.

[7] I. Marzona, F. Avanzini, G. Lucisano, M. Tettamanti, M. Baviera,
A. Nicolucci, and M. C. Roncaglioni, ''Are all people with diabetes and cardiovascular risk factors or microvascular complications at very high risk? Findings from the risk and prevention study,'' Acta Diabetolog., vol. 54, no. 2, pp. 123–131, Feb. 2017.

[8] S. Basu, J. B. Sussman, S. A. Berkowitz, R. A. Hayward, and J. S. Yudkin, ''Development and validation of risk equations for complications of type 2 diabetes (RECODe) using individual participant data from randomized trials,'' Lancet Diabetes Endocrinol.,vol. 5, no. 10, pp. 788–798, Oct. 2017.

[9] E. B. Schroeder, S. Xu, G. K. Goodrich, G. A. Nichols, P. J. O'Connor, and J. F. Steiner, ''Predicting the 6-month risk of severe hypoglycemia among adults with diabetes: Development and external validation of a prediction model,'' J. Diabetes Complications, vol. 31,no. 7, pp. 1158–1163, Jul. 2017