ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25apr1043

Machine Learning Model for Prediction of Prediabetes Among Adults in Nigeria and Ghana

Adedoyin O. Taiwo[1]

[1] Department of Epidemiology and Medical Statistics, Faculty of Public Health, College of Medicine, University of Ibadan, Ibadan, Nigeria.

Publication Date: 2025/05/03

Abstract

- ***Introduction:***

Pre-diabetes is a significant metabolic disease that can have harmful effects on the body as a whole, with millions of cases in Africa. Early identification and treatment of pre-diabetes is necessary to decrease the risk of diabetes, as well as maintaining a healthy lifestyle. Machine learning, on the other hand, is a computational method for automated learning from data for accurate predictions. Deploying machine learning models for the prediction of health outcomes in clinical medicine (including oncology, cardiovascular diseases, and diabetes), is now gaining wave around the globe, however, there is no such model available for the prediction of pre-diabetes among Africans. Hence, there is a need for an Afrocentric model that identifies the risk of developing pre-diabetes among Africans.

- ***Objective:***

The aim of this study is to build such model that would help in predicting the outcome of Pre-Diabetes among adult Nigerians and Ghanaians for proper diagnosis and disease preventive measures.

- ***Methods:***

The data analysed in this research included 2463 participants from Nigeria and Ghana. Further Pre-processing of the data, which involved excluding those participants that are already diabetic" left this research with 2,016 research participants. The outcome variable is a recode of the Laboratory Fasting Blood Glucose variable where the participants with < 99mg/dl are normal, participants with Laboratory Fasting Blood Glucose between 100mg/dl and 125mg/dl are pre-diabetic, and participants with Laboratory Fasting Blood Glucose > 125mg/dl are diabetic. This study assessed five different supervised machine learning predictive models, including Support vector machine (SVM), k-NN, Naïve Bayes, Random Forest, Decision Tree Classifier and Logistic Regression to predict diagnostic outcomes for pre-diabetes. The performance of all the five distinct models were assessed using precision, recall, area under curve (AUC) and F1 score.

- ***Results:***

The result of this study also showed that 10% of the study participants considered are prediabetic. Family history (OR = 41.50), Hypertension Status (OR = 1.53), Tobacco Use (OR = 1.05), Alcohol Use (OR = 1.01), BMI (OR = 1.04), and Obesity (OR = 1.28) are factors that increase prediabetes outcome. The results of our feature selection methods showed that Domicile, Alcohol Use, Family History, Tobacco Use, Dyslipidemia, Body Mass Index (BMI), Age, Obesity, Blood Pressure, Hypertension Status, Country, Gender contributed more to the prediction of prediabetes outcome. The areas under curve and accuracy results for all models showed that Random Forest (0.90, 0.85), SVM (0.92, 0.86) and the logistic regression model (0.92, 0.86) performed best on classification accuracy.

- ***Conclusion:***

The study concluded that the Support Vector Machine (SVM) is the most efficient model in predicting prediabetes outcome. Hence, SVM can be integrated into medical devices and software applications to determine prediabetic outcome among Adults in Nigeria and Ghana. This study will also aid future researchers in selecting the most suitable predictive models for the implementation of community lifestyle programs aimed at reducing the prevalence of prediabetes.

Keywords*: Machine Learning, Prediabetes, Support vector machine, k-NN, Naïve Bayes, Random Forest, Decision Tree Classifier and Logistic Regression.*

How to cite: Adedoyin O. Taiwo (2025). Machine Learning Model for Prediction of Prediabetes Among Adults in Nigeria and Ghana. *International Journal of Innovative Science and Research Technology*, 10(4), 2367-2377 https://doi.org/10.38124/ijisrt/25apr1043

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25apr1043

I. INTRODUCTION

The attention of the global world is now increasingly drawn to the utilization of prognostic and diagnostic prediction models in several domains of Healthcare outcome research and clinical medicines including cancers, cardiovascular diseases, hypertension, prediabetes, and diabetes (Lynam et al., 2020). These models are experiencing a growing utilization as web-based calculators and medical applications for smartphones, with a significant number of them being integrated into clinical recommendations (Wessler et al., 2016). The availability of these models in such easily accessible devices have increased the efficiency of disease surveillance, disease prediction, targeted clinical interventions, and evidence-based research. More so, with the advancement in knowledge, the global science community has continually developed different approaches and methodologies to the development and deployment of these models. Historically, traditional statistical models like logistic regression have been often utilized. However, there has been a growing inclination towards the utilization of more robust machine learning techniques to enhance prognostic and diagnostic precision in the field of clinical research (Borson et al., 2020) With many evidence of its use, Machine learning has brought about unprecedented advancement in prognostic and diagnostic predictions.(Kavakiotis et al., 2017; Ogallo et al., 2020)

Additionally, James He, (2014) described Machine learning (ML) as a discipline within the realm of data science that focuses on the advancement of algorithms and methodologies enabling computers (referred to as machines) to dynamically adjust, acquire knowledge, and exhibit intelligence by leveraging on real world data. Learning in this case refers to the process by which a system acquires the ability to recognize and comprehend the input data, enabling it to generate decisions and predictions based on this acquired knowledge. These algorithms are built is such a way that enables them to effectively handle large volumes of data, including medical imaging, biobank information, real-time medical readings, and electronic health care records. The process of learning and relearning from the data is called training, the optimization of this process produces a model that explains the underlying factors and nuances in the data with a high level of accuracy.

Prediabetes is a medical disease which is characterized by elevated blood glucose levels that exceed the normal range, however, do not meet the diagnostic criteria for diabetes mellitus. According to Martins et al. (2017), Prediabetes is a transitional phase that occurs between normal glucose tolerance (NGT) and the development of overt type 2 diabetes mellitus. It is characterized by elevated blood glucose levels that beyond the normal range but do not meet the diagnostic criteria for diabetes mellitus. Therefore, it encompasses two distinct cohorts: those with impaired glucose tolerance (IGT) and individuals with impaired fasting glucose (IFG). As per the guidelines provided by the American Diabetes Association, the condition known as prediabetes is identified through specific diagnostic criteria. These criteria include a fasting blood glucose level ranging from 100 to 125 mg/dl (5.6 to 6.9 mmol/L), referred to as impaired fasting glucose (IFG). Additionally, a blood glucose level ranging from 140 to 199 mg/dl (7.8 to 11.0 mmol/L) two hours following an oral glucose tolerance test (OGTT), known as impaired glucose tolerance (IGT), or a HbA1c level ranging from 5.7% to 6.4% can also indicate the presence of prediabetes. Prediabetes is predominantly characterized by the absence of noticeable symptoms, making its identification primarily reliant on normal screening procedures conducted in seemingly healthy persons. There are three screening procedures employed for the detection of prediabetes, including the assessment of fasting blood glucose (FBG) levels, the administration of a two-hour oral glucose tolerance test (OGTT), and the measurement of HbA1c levels. The oral glucose tolerance test (OGTT) is widely recognized as a reliable measure for assessing the risk of developing diabetes and is considered the most accurate method for identifying individuals with prediabetes (Martins et al., 2017). The objective of this research is to construct a machine learning framework that can accurately forecast the occurrence of prediabetes in the adult population residing in West Africa.

II. METHODOLOGY

The data used for this research included 2463 participants from Nigeria and Ghana. Further Preprocessing of the data, which involved excluding those participants that are already diabetic" left this research with 2,016 research participants. Furthermore, the explanatory variables considered for this research are selected because of findings from past literatures. The predictors that were considered, as justified to be risk factors for diabetes from previous literatures, are Blood Pressure, Body Mass Index, Family History of diabetes, Age, Gender, Tobacco use, Dyslipidaemia status, Hypertension status, Obesity status, Alcohol use, Domicile, Sleep Quality, Income. The outcome variable is a recode of the Laboratory Fasting Blood Glucose variable where the participants with < 99mg/dl are considered to be normal, while the participants with Laboratory Fasting Blood Glucose between 100mg/dl and 125mg/dl are considered to be pre-diabetic, and the participants with Laboratory Fasting Blood Glucose > 125mg/dl are considered to be diabetic.

- *The Python Programming Language (Spyder Version 5) was used to analyze the data. Firstly,it was used for descriptive analysis (to summarize the study participants' background characteristics). It was also used for the Machine Learning Algorithms listed below:*

- Logistic Regression
- K Nearest Neighbor
- Decision Tree
- Random Forest
- Bayesian Model
- Support Vector Machine

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25apr1043

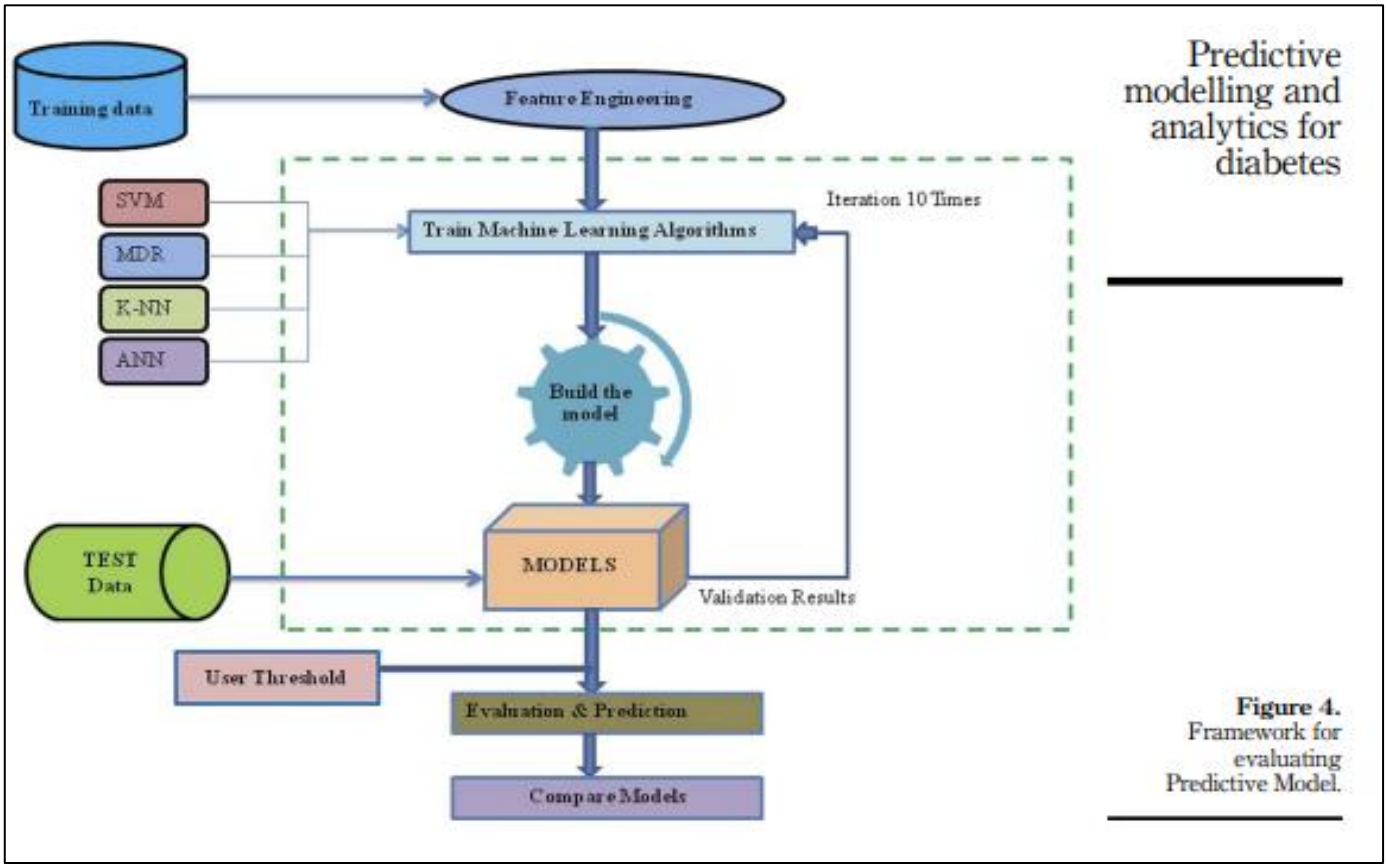

Fig 1 Implementation Structure of the Models

➢ *The implementation of the machine learning models was done in 7 steps:*

- Import Relevant Libraries
- Read in Data, Perform Exploratory Data Analysis (EDA).
- Feature Selection
- Create Feature (X) And Target (Y) Dataset
- Split Data into 70:30 Ratio, Where 70 (Training Set), While 30 (Validation Test).
- Train The Model.
- Compute The Performance Metrics.

The performance metrics to be used in the evaluating the validity of the predictive model are: Accuracy Score, Precision Score, Recall Score, Area under ROC with 95% confidence interval, Sensitivity, Specificity.

III. RESULT

The demographic characteristics of the 2,016 participants in this research shows that the average age of participants was found to be about 58 $(\pm 0.54\ years)$, the variation in Age was found to be about 14 years. The oldest participant is 100 years old, while the youngest participant is 20 years. The age distribution shows that the participants in this research are adults. The average Body Mass Index (BMI) of the 2,016 participants is about 27 $(\pm 0.23\ kg/m^2)$. The variability in BMI is found to be around 5.8. The highest observed BMI is about 62, while the lowest observed BMI is abourt12. The average blood pressure of participants is about 83 mm/hg, the highest blood pressure observed is about 159 mm/hg while the lowest blood pressure observed is about 40 mm/hg.

Also, the proportion of Male (52%) is more than their female (48%) counterparts. It was also found that the proportion of participants living in Urban areas (623%) is more than those living in Rural (13%) and Semi Urban (24%). It was also found that most of the respondents have Poor sleep quality (96%), with a little proportion (4%) having good quality sleep. Most of the respondents are non-obese (97%), with a little percentage (3%) being obese.

It was further found that the number of participants in this research that are hypertensive (1520, 60%) is more than the participants that are non-hypertensive (943, 40%). 67% of the participants have dyslipidemia, while 33% does not have. 13% of the participants has Family History of Diabetes, while 87% do not have Family History of diabetes. Also, it was found that 79% of the participants currently use Alcohol, and 92% currently use tobacco. It was also found that 10% of the participants are Prediabetic.

IV. FEATURE SELECTION

The lasso CV and Random Forest algorithm was used to reduce the dimensionality of the dataset, to eliminate non-relevant variables. The most important variables that were used for the training of the algorithms are Domicile, Alcohol Use, Family History, Tobacco Use, Dyslipidemia, Body Mass Index (BMI), Age, Obesity, Blood Pressure, Hypertension Status, Country, and Gender

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25apr1043

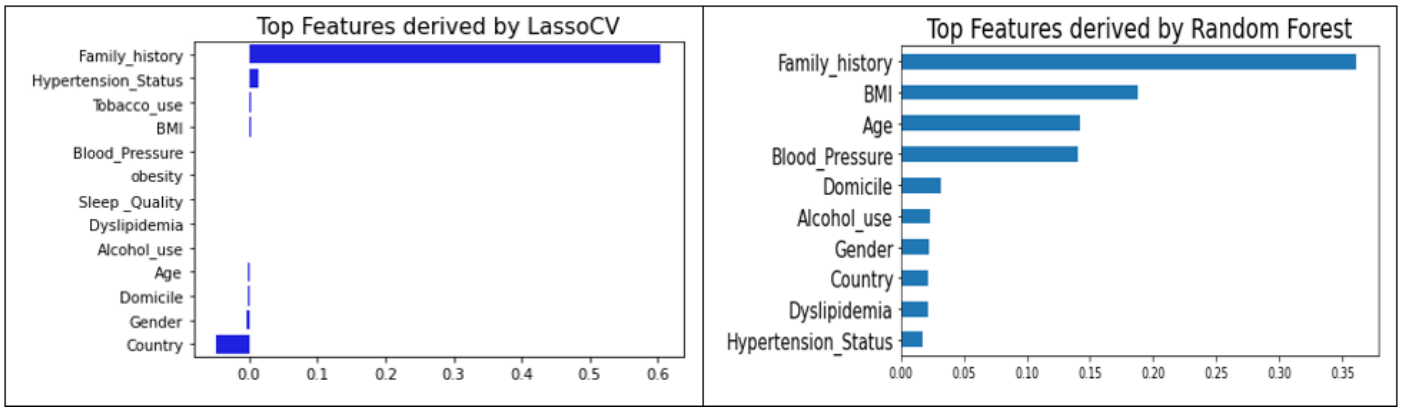

Fig 2 Feature Selection Chart

➢ *Performance of variables*

The Odds Ratio table shows that Family history (OR = 41.50), Hypertension Status (OR = 1.53), Tobacco Use (OR = 1.05), Alcohol Use (OR = 1.01), and Obesity (OR = 1.28) are factors that increase prediabetes outcome. It also shows that the Family history (OR = 41.50, CI = 28.69 - 60.03), country of residence (OR = 0.43, CI = 0.29 - 0.64) and BMI (OR = 1.04, CI = 1.01 - 1.07) are significantly associated with the outcome of prediabetes.

Table 1 Performance of Variables

Prediabetes status	odds ratio	Std. Err.	p-value	[95% conf. Interval]
Age	0.99	0.01	0.37	0.98 - 1.01
Domicile	0.93	0.12	0.58	0.72 - 1.20
Gender	0.86	0.16	0.42	0.59 - 1.25
BMI	1.04	0.02	0.03	1.01 - 1.07
Blood pressure	1.00	0.01	0.87	0.99 - 1.02
Obesity	1.28	0.53	0.55	0.57 - 2.90
Hypertension status	1.53	0.36	0.07	0.97 - 2.41
Dyslipidemia	0.98	0.19	0.93	0.68 - 1.43
Family history	41.50	7.82	0.00	28.69 - 60.03
Tobacco use	1.05	0.18	0.77	0.75 - 1.48
Alcohol use	1.01	0.21	0.98	0.67 - 1.52
Country	0.43	0.09	0.00	0.29 - 0.64

➢ *Model selection*

The reduced dataset was trained using five different Machine Learning Algorithms. Which are

- Logistic Regression
- K Nearest Neighbor
- Decision Tree
- Random Forest
- Bayesian Model
- Support Vector Machine

➢ *Evaluation of the Trained Models*

To evaluate the trained models the study considered the Area under the Receiver Operating curve for each model and their corresponding confusion matrix. This was deployed to measure the models' accuracy in predicting pre-diabetes outcome among adults in Nigeria and Ghana.

➢ *Receivers operating Curve of the Logistics Regression Model*

The logistic Regression model has an Area under Curve (AUC) of its Receiver Operating Curve to be 86%.

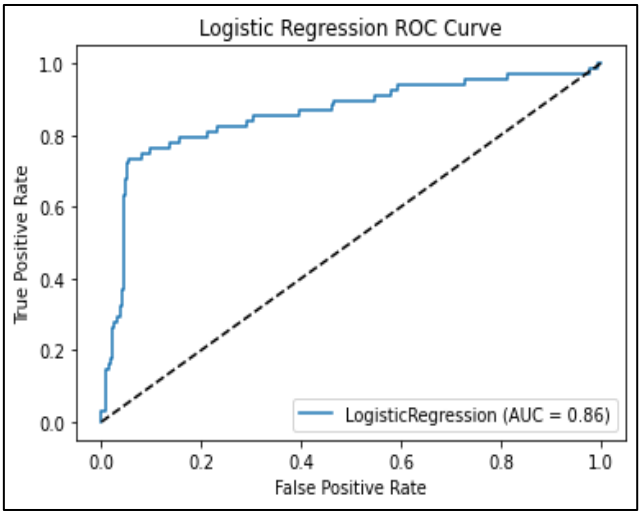

Fig 3 Logistics Regression ROC Curve

➢ *The Confusion Matrix for The Logistic Regression Model*

The confusion matrix as shown in fig 4 below shows that the logistics regression model has an accuracy of 92% in the prediction of prediabetes outcome.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25apr1043

Table 2 Confusion Matrix for Logistics Regression

		Actual	
		Positive	Negative
Predicted	Positive	415	21
	Negative	25	43

➢ *Receivers operating Curve of the KNN Model*

The K Nearest Neighbors model has an Area under Curve (AUC) of its Receiver Operating Curve to be 86%.

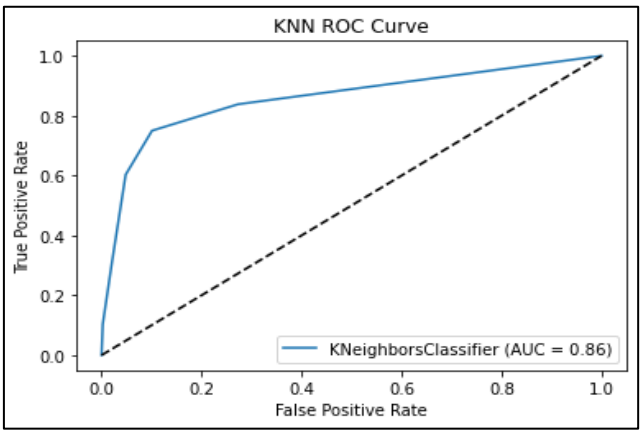

Fig 4 KNN ROC Curve

➢ *The Confusion Matrix for The KNN Model*

The confusion matrix as shown in figure 6 below shows that the KNN model has an accuracy of 90% in the prediction of prediabetes outcome.

Table 3 Confusion Matrix for KNN

		Actual	
		Positive	Negative
Predicted	Positive	415	21
	Negative	27	41

➢ *Receivers Operating Curve of the Decision Tree Model*

The Decision Tree model has an Area under Curve (AUC) of its Receiver Operating Curve to be 67%.

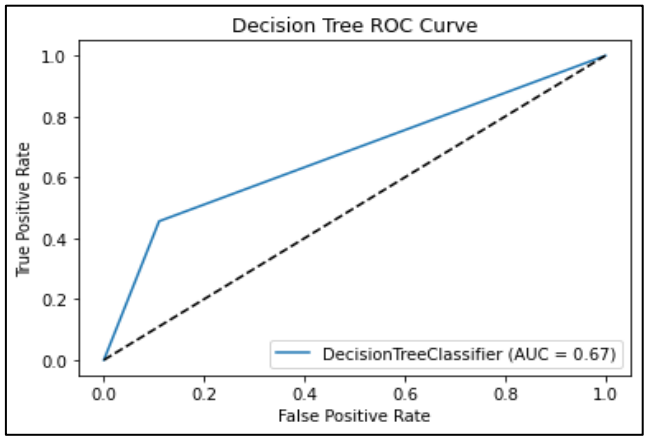

Fig 5 Decision Tree ROC Curve

➢ *The Confusion Matrix for the Decision Tree Model*

The confusion matrix as shown in figure 8 below shows that the Decision tree model has an accuracy of 86% in the prediction of prediabetes outcome.

Table 4 Decision Tree Confusion Matrix

		Actual	
		Positive	Negative
Predicted	Positive	388	48
	Negative	37	31

➢ *Receivers Operating Curve of the Random Forest Model*

The Random Forest model has an Area under Curve (AUC) of its Receiver Operating Curve to be 85%.

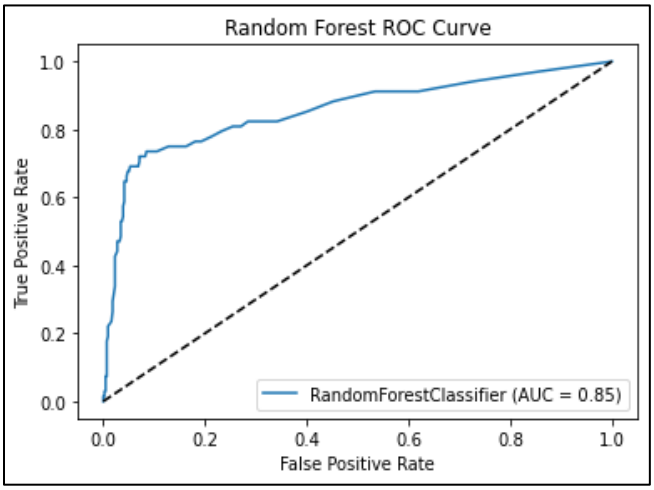

Fig 6 Random Forest ROC Curve

➢ *The confusion Matrix for the Random Forest Model*

The confusion matrix as shown in figure 10 below shows that the Random Forest model has an accuracy of 90% in the prediction of prediabetes outcome.

Table 5 Random Forest Confusion Matrix

		Actual	
		Positive	Negative
Predicted	Positive	420	16
	Negative	32	36

➢ *Receivers Operating Curve of the Gaussian Naïve Bayes Model*

The Gaussian Naïve Bayes model has an Area under Curve (AUC) of its Receiver Operating Curve to be 83%.

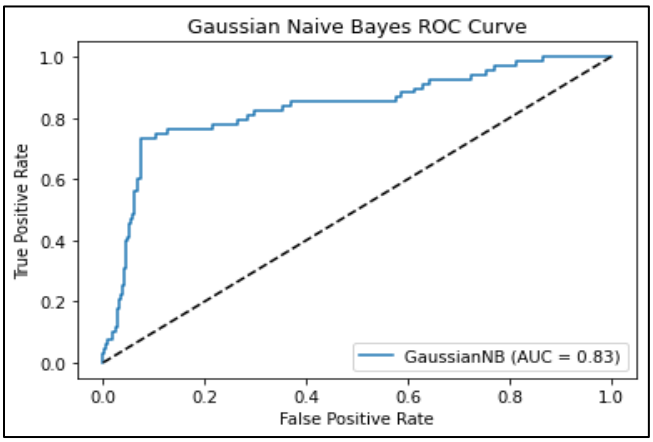

Fig 7 Gaussian Naïve Bayes ROC Curve

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25apr1043

➢ *The Confusion Matrix for the Gaussian Naïve Bayes Model*

The confusion matrix as shown in figure 12 below shows that the Gaussian Naïve Bayes model has an accuracy of 90% in the prediction of prediabetes outcome.

Table 6 Gaussian Naïve Bayes Confusion Matrix

		Actual	
		Positive	Negative
Predicted	Positive	415	21
	Negative	30	38

➢ *Receivers operating Curve of the Bernoulli Bayes Model*

The Bernoulli Naïve Bayes model has an Area under Curve (AUC) of its Receiver Operating Curve to be 84%.

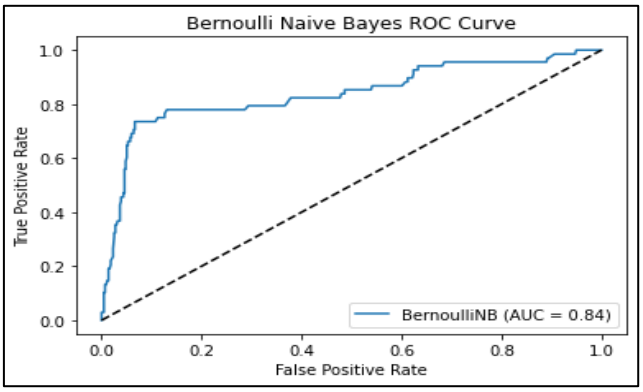

Fig 8 Bernoulli Bayes ROC Curve

➢ *The confusion Matrix for the Bernoulli Naïve Bayes Model*

The confusion matrix as shown in figure 14 below shows that the Bernoulli Naïve Bayes model has an accuracy of 90% in the prediction of prediabetes outcome.

Table 7 Bernoulli Bayes Confusion Matrix

		Actual	
		Positive	Negative
Predicted	Positive	415	21
	Negative	30	38

➢ *Receivers Operating Curve of the Linear Support Vector Machine Model*

The Linear Support Vector Machine model has an Area under Curve (AUC) of its Receiver Operating Curve to be 83%.

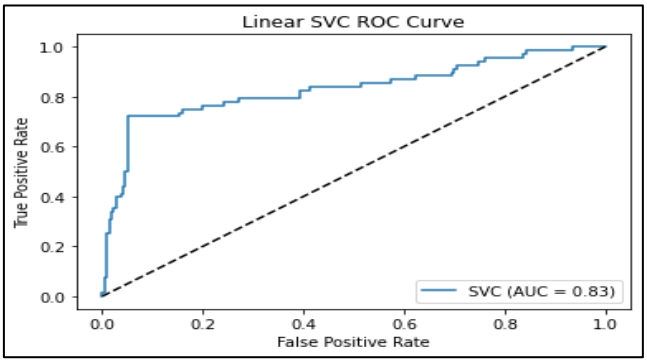

Fig 9 Linear SVC ROC Curve

➢ *The Confusion Matrix for The Linear Svc Model*

The confusion matrix as shown in figure 16 below shows that the Linear SVC model has an accuracy of 92% in the prediction of prediabetes outcome.

Table 8 Linear Svc Confusion Matrix

		Actual	
		Positive	Negative
Predicted	Positive	416	20
	Negative	24	44

➢ *Receivers Operating Curve of The Support Vector Machine Model*

The Gaussian Support Vector Machine model has an Area under Curve (AUC) of its Receiver Operating Curve to be 86%.

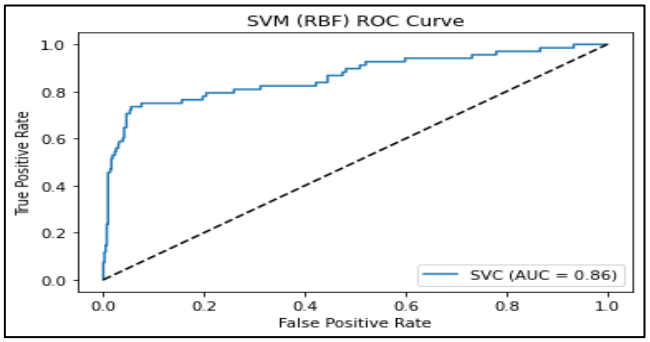

Fig 10 SVC (RBF) ROC Curve

➢ *The Confusion Matrix for The Svm Model*

The confusion matrix as shown in figure 18 below shows that the SVM model has an accuracy of 92% in the prediction of prediabetes outcome.

Table 9 SVC (RBF) Confusion Matrix

		Actual	
		Positive	Negative
Predicted	Positive	416	20
	Negative	24	44

➢ *Performance Criterion After Hyper Parameter Tuning*

Table 2 shows the performance of all the models, after tuning their various Hyper-Parameters, where they are evaluated upon parameters like precision, recall, area under curve (AUC) and F1score. From table 2, we found that the accuracy of Logistic Regression, KNN, Decision Tree, Random Forest, Gaussian Naïve Bayes, Bernoulli Naïve Bayes, Linear SVC, and SVC are 0.92, 0.90, 0.67, 0.90, 0.90, 0.90, 0.92, and 0.92 respectively. It was also found that the Area under the ROC curve for the models under consideration (Logistic Regression, KNN, Decision Tree, Random Forest, Gaussian Naïve Bayes, Bernoulli Naïve Bayes, Linear SVC, and SVC) are 0.86, 0.86, 0.86, 0.85, 0.83, 0.84, 0.83, and 0.86 respectively. The sensitivity/precision of the models are 0.67, 0.66, 0.42, 0.69, 0.56, 0.68, 0.67, and 0.68 respectively. The log-loss estimate for the models are 3.22, 3.15, 5.14, 3.22, 3.50, 3.56, 2.88, and 2.88 respectively. Therefore, most efficient models for classification of the pre-diabetic cases are the Logistic Regression model, Random Forest Model and the Gaussian SVC model.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25apr1043

Table 10 Performance criterion Table after hyper parameter tuning

Model	accuracy	precision	Recall	f1score	rocauc	logloss
Logistic Regression	0.92	0.67	0.63	0.66	0.86	3.22
KNN	0.90	0.66	0.53	0.58	0.86	3.15
Decision Tree	0.86	0.42	0.46	0.44	0.67	5.14
Random Forest	0.90	0.66	0.50	0.56	0.85	3.22
Gaussian NB	0.90	0.56	0.69	0.61	0.83	3.50
Bernoulli NB	0.90	0.69	0.60	0.63	0.84	3.56
Linear SVC	0.92	0.67	0.67	0.67	0.83	2.88
SVC (RBF)	0.92	0.66	0.62	0.64	0.86	2.88

➢ *Predictive Performance of the Most Efficient Models:*

The parameters of the most efficient model are now tuned to get the best estimates and prediction accuracy for each model.

➢ *The Performance of Logistic Regression Model*

With an accuracy of 92%, the performance criteria show that the models will predict a prediabetic patient positive around 67% of the time. Also, it shows that the model will at a probability of 95% predict a non-prediabetic adult negative. It also shows that if the prediction is positive, then there is about 57% chance that the adult is prediabetic. Also, if the prediction is negative, then there is about 96% chance that the adult is non-prediabetic.

Table 11 The Performance of Logistic Regression Model

Statistic	Value	95% CI
Sensitivity	69.19%	54.31% to 78.41%
Specificity	94.32%	91.73% to 96.29%
Positive Likelihood Ratio	11.83	7.79 to 17.95
Negative Likelihood Ratio	0.35	0.24 to 0.49
Disease prevalence (*)	10.00%	
Positive Predictive Value (*)	56.78%	46.40% to 66.61%
Negative Predictive Value (*)	96.28%	94.79% to 97.35%

➢ *The Performance of Random Forest Model*

With an accuracy of 90%, the performance criteria show that the models will predict a prediabetic patient positive around 69% of the time. Also, it shows that the model will at a probability of 94% predict an a non-prediabetic adult negative. It also shows that if the prediction is positive, then there is about 55% chance that the adult is prediabetic. Also, if the prediction is negative, then there is about 97% chance that the adult is non-prediabetic.

Table 12 The Performance of Random Forest Model

Statistic	Value	95% CI
Sensitivity	68.97%	55.46% to 80.46%
Specificity	93.72%	91.05% to 95.79%
Positive Likelihood Ratio	10.99	7.38 to 16.36
Negative Likelihood Ratio	0.33	0.23 to 0.49
Disease prevalence (*)	10.00%	
Positive Predictive Value (*)	54.97%	45.05% to 64.50%
Negative Predictive Value (*)	96.45%	94.87% to 97.56%

➢ *The Performance of SVC Model*

With an accuracy of 92%, the performance criteria show that the models will predict a prediabetic patient positive around 68% of the time. Also, it shows that the model will at a probability of 96% predict an a non-prediabetic adult negative. It also shows that if the prediction is positive, then there is about 63% chance that the adult is prediabetic. Also, if the prediction is negative, then there is about 96% chance that the adult is non-prediabetic.

Table 13 The Performance of SVC Model

Statistic	Value	95% CI
Sensitivity	68.06%	56.01% to 78.56%
Specificity	95.60%	93.22% to 97.33%
Positive Likelihood Ratio	15.47	9.70 to 24.69
Negative Likelihood Ratio	0.33	0.24 to 0.47
Disease prevalence (*)	10.00%	
Positive Predictive Value (*)	63.23%	51.87% to 73.29%
Negative Predictive Value (*)	96.42%	95.05% to 97.42%

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25apr1043

V. DISCUSSION

The timely identification of prediabetes can significantly contribute to enhancing patients' quality of life and increasing their life expectancy. Previous research has mostly focused on the application of supervised algorithms in the development of various models for diabetes detection. However, there is a noticeable gap in the literature about the exploration of Pre-Diabetes detection (de Silva et al., 2020).

Various classification approaches were implemented and tested using the Python Spyder Integrated Development Environment (IDE). The dataset was divided into two distinct subsets, namely the training set and the testing set. Our model was trained using 70% of the available data, while the remaining 30% was used for testing. This partitioning proportion was consistent for both the training and testing phases (Li et al., 2021). Five different models have been developed using supervised learning to predict prediabetic outcome in the adult population.

The Odds Ratio table shows that Family history (OR = 41.50), Hypertension Status (OR = 1.53), Tobacco Use (OR = 1.05), Alcohol Use (OR = 1.01), and Obesity (OR = 1.28) are factors that increase prediabetes outcome. It also shows that the Family history (OR = 41.50, CI = 28.69 - 60.03), country of residence (OR = 0.43, CI = 0.29 - 0.64) and BMI (OR = 1.04, CI = 1.01 - 1.07) are significantly associated with the outcome of prediabetes. This is in line with (Sangrós et al., 2018) that reported that obesity is a strong predictor of prediabetes, and (Hubbard et al., 2019) that reported that Hypertension status has a strong association with prediabetes status. It was found that the Family history has the strongest significant association with prediabetes outcome, which is consonance with the findings of (Wagner et al., 2013), who highlighted that Family History of diabetes is a very strong predictor for prediabetes.

The Random Forest and Lasso CV showed that Domicile, Alcohol Use, Family History, Tobacco Use, Dyslipidemia, Body Mass Index (BMI), Age, Obesity, Blood Pressure, Hypertension Status, Country, Gender contributed more to the prediction of prediabetes outcome among Nigeria and Ghanaian adults. This is consistent with the findings of (Liu et al., 2021) that maintained that blood pressure is associated with increased pre-diabetes among U.S. adolescents, and (Martins et al., 2017) who found that sex and positive family history of diabetes mellitus, alcohol intake are associated with increased prediabetic outcome.

The Lasso Cross Validation and the Random Forest, which has shown consistency in feature selection (Choi et al., 2014), model shows that Ethnicity, Fasting Glucose, Family History, Alcohol use, Body Mass Index, Age, Blood Pressure, Domicile, Dyslipidemia and Gender are the most important factors for the prediction of prediabetes outcomes among Nigerian and Ghanaian Adults.

For this study, linear kernel support vector machine (SVM-linear), k-NN, Naïve Bayes, Random Forest, Decision Tree Classifier and Logistic Regression were used for training the data, to predict diagnostic outcomes for pre-diabetes outcome for the Nigerian and Ghanaian adult population. The evaluation of the performance of the five distinct models involved the assessment of various metrics such as precision, recall, area under curve (AUC), and F1 score. These measures are widely recognized as standard performance criteria in the field (de Silva et al., 2020).

In order to mitigate the issues of overfitting and underfitting, researchers conducted a fivefold cross-validation procedure (Arowolo et al., 2022). The accuracy of our classifier refers to the frequency with which it correctly diagnoses whether a patient is pre-diabetic or not. Precision has been employed as a metric to assess the classifier's capacity to accurately predict positive cases of pre-diabetes. In our study, the metrics of recall or sensitivity are employed to determine the accuracy with which the classifier properly identifies the fraction of true positive instances of pre-diabetes. The utilization of specificity is employed to assess the classifier's capacity to accurately identify individuals who do not have pre-diabetes. The F1 score is derived from the weighted mean of precision and recall, so including both measures into a single metric. Classifiers with an F1 score close to 1 are referred regarded be the most optimal ones (Hinton & Sejnowski, 1999).

The Receiver Operating Characteristic (ROC) curve is a widely recognized technique utilized for visualizing the performance of a binary classifier system, as mentioned by Choi et al. (2014). The graph depicts the relationship between the true positive rate and the false positive rate, while the threshold for classifying observations into a certain category is adjusted. The range of the area under the curve (AUC) value for a classifier often falls within the interval of 0.5 to 1. Values below 0.50 in a given dataset imply an inability to differentiate between true and false due to randomness. An ideal classifier is characterized by a high value of the area under the curve (AUC), approaching 1.0. When the value approaches 0.5, it can be seen as being on par with random guessing, as noted by de Silva et al. (2020).

The study found that the performance of logistic regression, Random Forest and Support Vector Machine is better than the other machine learning algorithms to classify Prediabetes Cases. Discrimination was high for all models, while Support Vector Machine model showed the numerically highest discrimination in criteria measures. This is also in concordance with (Choi et al., 2014) that reported the SVM being the best model for prediabetes prediction.

One of the notable strengths of this study is the implementation of a systematic method to model comparison. This strategy involves utilizing the identical dataset for training all models. This is particularly important as the performance of models can vary across different contexts. By employing the same dataset, the study ensures the validity of model comparisons. The selection of tuning parameters has a significant impact on the model's performance. To enhance our models, we have conducted hyperparameter tuning using a well-established grid search methodology. We have conducted a comparative analysis of multiple machine

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25apr1043

learning algorithms, specifically chosen for their appropriateness within our specific context. The utilization of a limited set of eleven predictor variables mitigates the potential issue of overfitting. In the context of machine learning algorithms, it has been proposed that a minimum of 10 times the number of events per variable is necessary to attain reliable outcomes, in contrast to conventional statistical modeling. Understanding the effectiveness of machine learning models while utilizing an optimal number of predictors holds significant importance. It is conceivable that the inclusion of more variables or a larger sample size may enhance the discriminative capabilities of machine learning systems.

Nevertheless, the study possesses inherent limitations. One disadvantage of this study pertains to the sole reliance on evaluating the model's performance. When making a decision on the most suitable model, it is important to take into account the factors of implementation and interpretation for practical purposes.

VI. CONCLUSION

In summary, we compared five prediction models for prediabetes outcome using 13 risk factors. The results indicated that the Random Forest, SVM and the logistic regression model performed best on classification accuracy, with the Support Vector Machine (SVC) being the most efficient model in predicting prediabetes outcome. This is rightly in concordance with (Choi et al., 2014). The objective of this study is to provide guidance to future researchers in selecting the most effective predictive models for the implementation of community lifestyle interventions aimed at reducing the prevalence of prediabetes.

RECOMMENDATIONS

The huge burden of prediabetes and diabetes cases in Nigeria represents a unique set of problems and provides the us with a unique opportunity in terms of potential availability of data. Harnessing this data using electronic medical records, by all physicians, can put Nigeria at the forefront of research in this area. Application of AI/ML would provide insights to our problems as well as may help us to devise tailor-made solutions for adult Nigerians and Ghanaians.

This study hereby recommends that various EMR software be developed to facilitate availability of data and the models developed in this study be integrated into medical devices that will help in the prediction of prediabetes outcome.

REFERENCES

[1]. Ahmed, S. M., & al Mansour, M. (2017). A study on the prevalence of risk factors for diabetes and hypertension among school children in Majmaah, Kingdom of Saudi Arabia. Journal of Public Health Research, 6(2). https://doi.org/10.4081/jphr.2017.829

[2]. Anandkumar, A., Ge, R., Hsu, D., Kakade, S. M., & Telgarsky, M. (2012). Tensor decompositions for learning latent variable models. http://arxiv.org/abs/1210.7559

[3]. Arowolo, M. O., Ogundokun, R. O., Misra, S., Kadri, A. F., & Aduragba, T. O. (2022). Machine Learning Approach Using KPCA-SVMs for Predicting COVID-19. In EAI/Springer Innovations in Communication and Computing. https://doi.org/10.1007/978-3-030-72752-9_10

[4]. Bashir, M. I., Wani, A. I., & Masoodi, S. R. (2011). Prediabetes. JMS SKIMS, 14(1), 4–10. https://doi.org/10.33883/JMS.V14I1.62

[5]. Borson, N. S., Kabir, M. R., Zamal, Z., & Rahman, R. M. (2020). Correlation analysis of demographic factors on low birth weight and prediction modeling using machine learning techniques. Proceedings of the World Conference on Smart Trends in Systems, Security and Sustainability, WS4 2020. https://doi.org/10.1109/WorldS450073.2020.9210338

[6]. CDC US departement of Health and Human Services. (2020). National Diabetes Statistics Report, 2020 Estimates of Diabetes and Its Burden in the United States. In National Diabetes Statistics Report.

[7]. Choi, S. B., Kim, W. J., Yoo, T. K., Park, J. S., Chung, J. W., Lee, Y. H., Kang, E. S., & Kim, D. W. (2014). Screening for prediabetes using machine learning models. Computational and Mathematical Methods in Medicine, 2014. https://doi.org/10.1155/2014/618976

[8]. Comesaña-Campos, A., & Bouza-Rodríguez, J. B. (2016). An application of Hebbian learning in the design process decision-making. Journal of Intelligent Manufacturing, 27(3), 487–506. https://doi.org/10.1007/s10845-014-0881-z

[9]. de Silva, K., Jönsson, D., & Demmer, R. T. (2020). A combined strategy of feature selection and machine learning to identify predictors of prediabetes. Journal of the American Medical Informatics Association, 27(3). https://doi.org/10.1093/jamia/ocz204

[10]. Duun-Henriksen, A. K., Schmidt, S., Røge, R. M., Møller, J. B., Nørgaard, K., Jørgensen, J. B., & Madsen, H. (2013). Model identification using stochastic differential equation grey-box models in diabetes. Journal of Diabetes Science and Technology, 7(2). https://doi.org/10.1177/193229681300700220

[11]. Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., Kim, R., Raman, R., Nelson, P. C., Mega, J. L., & Webster, D. R. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA - Journal of the American Medical Association, 316(22). https://doi.org/10.1001/jama.2016.17216

[12]. Hinton, G. E., & Sejnowski, T. J. (Terrence J. (1999). Unsupervised Learning: Foundations of Neural Computation. MIT Press.

[13]. Hubbard, D., Colantonio, L. D., Tanner, R. M., Carson, A. P., Sakhuja, S., Jaeger, B. C., Carey, R. M., Cohen, L. P., Shimbo, D., Butler, M., Bertoni, A. G., Langford, A. T., Booth, J. N., Kalinowski, J., & Muntner, P. (2019). Prediabetes and risk for

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25apr1043

cardiovascular disease by hypertension status in black adults: The Jackson Heart Study. Diabetes Care, 42(12). https://doi.org/10.2337/dc19-1074

[14]. Ismail, L., Materwala, H., & al Kaabi, J. (2021). Association of risk factors with type 2 diabetes: A systematic review. In Computational and Structural Biotechnology Journal (Vol. 19). https://doi.org/10.1016/j.csbj.2021.03.003

[15]. Itohan, A. M., Khalid, F.-S., Badmos, A. O., & Alaran, A. J. (2021). Covid 19 Emphasizes the Need to Build Research Capacity in Africa. ADVANCES IN MEDICAL, DENTAL, 9-10.

[16]. James He, X. (2014). Business Intelligence and Big Data Analytics: An Overview. In Communications of the IIMA (Vol. 14). https://scholarworks.lib.csusb.edu/ciimaAvailableat:https://scholarworks.lib.csusb.edu/ciima/vol14/iss3/1

[17]. Jonathan, R. G., Ciaran, L. M., & Saurabh, J. (2020). Improving the accuracy of medical diagnosis with causal machine learning. Nature Communications, 3923.

[18]. Murray, S. (2019, April 19). Common causes of our diagnostic errors. From HCP live: https://www.hcplive.com/view/common-causes-of-our-diagnostic-errors

[19]. Kautzky-Willer, A., Harreiter, J., & Pacini, G. (2016). Sex and gender differences in risk, pathophysiology and complications of type 2 diabetes mellitus. In Endocrine Reviews (Vol. 37, Issue 3). https://doi.org/10.1210/er.2015-1137

[20]. Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine Learning and Data Mining Methods in Diabetes Research. In Computational and Structural Biotechnology Journal (Vol. 15). https://doi.org/10.1016/j.csbj.2016.12.005

[21]. Kengne, A. P., Amoah, A. G. B., & Mbanya, J. C. (2005). Cardiovascular complications of diabetes mellitus in sub-Saharan Africa. In Circulation (Vol. 112, Issue 23). https://doi.org/10.1161/CIRCULATIONAHA.105.544312

[22]. Lal, B. S. (2016). Diabetes : Causes, Symptoms and Treatment. Public Health Environment and Social Issue in India, January.

[23]. Liu, S., Gao, Y., Shen, Y., Zhang, M., Li, J., & Sun, P. (2019). Application of three statistical models for predicting the risk of diabetes. BMC Endocrine Disorders, 19(1). https://doi.org/10.1186/s12902-019-0456-2

[24]. Liu, C., Wu, S., & Pan, X. (2021). Clustering of cardio-metabolic risk factors and pre-diabetes among U.S. adolescents. Scientific Reports, 11(1). https://doi.org/10.1038/s41598-021-84128-6

[25]. Li, J., Yuan, P., Hu, X., Huang, J., Cui, L., Cui, J., Ma, X., Jiang, T., Yao, X., Li, J., Shi, Y., Bi, Z., Wang, Y., Fu, H., Wang, J., Lin, Y., Pai, C. H., Guo, X., Zhou, C., … Xu, J. (2021). A tongue features fusion approach to predicting prediabetes and diabetes with machine learning. Journal of Biomedical Informatics, 115. https://doi.org/10.1016/j.jbi.2021.103693

[26]. Lynam, A. L., Dennis, J. M., Owen, K. R., Oram, R. A., Jones, A. G., Shields, B. M., & Ferrat, L. A. (2020). Logistic regression has similar performance to optimised machine learning algorithms in a clinical setting: application to the discrimination between type 1 and type 2 diabetes in young adults. Diagnostic and Prognostic Research, 4(1). https://doi.org/10.1186/s41512-020-00075-2

[27]. Martins, S. O., Folasire, O. F., & Irabor, A. E. (2017). PREVALENCE AND PREDICTORS OF PREDIABETES AMONG ADMINISTRATIVE STAFF OF A TERTIARY HEALTH CENTRE, SOUTHWESTERN NIGERIA. Annals of Ibadan Postgraduate Medicine, 15(2).

[28]. Ogallo, W., Speakman, S., Akinwande, V., Varshney, K. R., Walcott-Bryant, A., Wayua, C., Weldemariam, K., Mershon, C. H., & Orobaton, N. (2020). Identifying Factors Associated with Neonatal Mortality in Sub-Saharan Africa using Machine Learning. AMIA ... Annual Symposium Proceedings. AMIA Symposium, 2020.

[29]. Ogbera, A. O. (2014). Diabetes mellitus in Nigeria: The past, present and future. World Journal of Diabetes, 5(6), 905. https://doi.org/10.4239/wjd.v5.i6.905

[30]. Oguejiofor, O. (2014). Diabetes in Nigeria: Impact, Challenges, Future Directions. Endocrinology & Metabolic Syndrome, 03(02). https://doi.org/10.4172/2161-1017.1000130

[31]. Okoronkwo, I. L., Ekpemiro, J. N., Okwor, E. U., Okpala, P. U., & Adeyemo, F. O. (2015). Economic burden and catastrophic cost among people living with type2 diabetes mellitus attending a tertiary health institution in south-east zone, Nigeria Endocrine Disorders. BMC Research Notes, 8(1). https://doi.org/10.1186/s13104-015-1489-x

[32]. Olesen, S. S., Poulsen, J. L., Novovic, S., Nøjgaard, C., Kalaitzakis, E., Jensen, N. M., Engjom, T., Tjora, E., Waage, A., Hauge, T., Haas, S. L., Vujasinovic, M., Barauskas, G., Pukitis, A., Ozola-Zālīte, I., Okhlobystin, A., Parhiala, M., Laukkarinen, J., & Drewes, A. M. (2020). Multiple risk factors for diabetes mellitus in patients with chronic pancreatitis: A multicentre study of 1117 cases. United European Gastroenterology Journal, 8(4). https://doi.org/10.1177/2050640620901973

[33]. Prediabetes | JMS SKIMS. (n.d.). Retrieved October 17, 2021, from http://www.jmsskims.org/index.php/jms/article/view/62

[34]. Sangrós, F. J., Torrecilla, J., Giráldez-García, C., Carrillo, L., Mancera, J., Mur, T., Franch, J., Díez, J., Goday, A., Serrano, R., García-Soidán, F. J., Cuatrecasas, G., Igual, D., Moreno, A., Millaruelo, J. M., Carramiñana, F., Ruiz, M. A., Pérez, F. C., Iriarte, Y., … Regidor, E. (2018). Association of General and Abdominal Obesity With Hypertension, Dyslipidemia and Prediabetes in the PREDAPS Study. Revista Espanola de Cardiologia, 71(3). https://doi.org/10.1016/j.recesp.2017.04.010

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25apr1043

[35]. Sarfo, F. S., Ovbiagele, B., Gebregziabher, M., Akpa, O., Akpalu, A., Wahab, K., Ogbole, G., Akinyemi, R., Obiako, R., Komolafe, M., Owolabi, L., Lackland, D., Arnett, D., Tiwari, H., Markus, H. S., Akinyemi, J., Oguntade, A., Fawale, B., Adeoye, A., … Owolabi, M. (2020). Unraveling the risk factors for spontaneous intracerebral hemorrhage among West Africans. Neurology, 94(10). https://doi.org/10.1212/WNL.0000000000009056

[36]. Singla, R., Singla, A., Gupta, Y., & Kalra, S. (2019). Artificial intelligence/machine learning in diabetes care. Indian Journal of Endocrinology and Metabolism, 23(4). https://doi.org/10.4103/ijem.IJEM_228_19

[37]. Suleiman, D. (2016). The persistent problem of diagnostic errors. Annals of Nigerian Medicine, 10(1). https://doi.org/10.4103/0331-3131.189800

[38]. Tripathy, J. P. (2018). Burden and risk factors of diabetes and hyperglycemia in india: Findings from the Global Burden of Disease Study 2016. Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy, 11. https://doi.org/10.2147/DMSO.S157376

[39]. Uloko, A. E., Musa, B. M., Ramalan, M. A., Gezawa, I. D., Puepet, F. H., Uloko, A. T., Borodo, M. M., & Sada, K. B. (2018). Prevalence and Risk Factors for Diabetes Mellitus in Nigeria: A Systematic Review and Meta-Analysis. Diabetes Therapy, 9(3). https://doi.org/10.1007/s13300-018-0441-1

[40]. Wagner, R., Thorand, B., Osterhoff, M. A., Müller, G., Böhm, A., Meisinger, C., Kowall, B., Rathmann, W., Kronenberg, F., Staiger, H., Stefan, N., Roden, M., Schwarz, P. E., Pfeiffer, A. F., Häring, H. U., & Fritsche, A. (2013). Family history of diabetes is associated with higher risk for prediabetes: A multicentre analysis from the German Center for Diabetes Research. Diabetologia, 56(10). https://doi.org/10.1007/s00125-013-3002-1

[41]. WHO. (2020). WHO Health 2020 pdf. From WHO Health 2020: https://www.who.int/workforcealliance/knowledge/resources/Health2020_long.pdf

[42]. Zand, A., Ibrahim, K., & Patham, B. (2018). Prediabetes: Why Should We Care? In Methodist DeBakey cardiovascular journal (Vol. 14, Issue 4). https://doi.org/10.14797/mdcj-14-4-289