Journal Homepage: - www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)

Article DOI: 10.21474/IJAR01/15940
DOI URL: http://dx.doi.org/10.21474/IJAR01/15940

RESEARCH ARTICLE

PREDICTIVE ANALYSIS OF DIABETES WITHOUT DATA PRE-PROCESSING VIA THE EVALUATION OF TREE ALGORITHMS

Silue Kolo[1], Johnson Grace Y. Edwige[1], Konan K. Hyacinthe[1], Asseu Olivier[1] and Bourget Daniel[2]

1. LASTIC, ESATIC, Abidjan, Côte d'Ivoire.
2. Lab-STICC, IMT-Atlantique, Brest, France.

Manuscript Info

Manuscript History
Received: 28 October 2022
Final Accepted: 30 November 2022
Published: December 2022

Key words:-
XGBoost, Decision Tree, Light GBM, Diabetes Prediction, PIMA Indians, Machine Learning, Cross Validation (k=5, 7, 12), Without Data Preprocessing

Abstract

Diabetes is a common disease, incurable and fatal in its complication phases. Its management, like many other metabolic diseases, remains a scientific challenge. Mathematical approaches have been used to understand this scourge and artificial intelligence is used to model its prediction. In general, the effectiveness and efficiency of an artificial intelligence solution depends on the nature and characteristics of the data and the performance of the learning methods. Hence the interest in the quality of the data and the performance of the methods used to model such a task. In order to find a suitable artificial intelligence model for diabetes prediction, several studies have used methods from different techniques. Thus, diabetes prediction has been addressed using machine learning methods, neural networks, deep learning, Bayesian naive classification, K-nearest neighbors and machine vector support. In order to compare the performance to determine the best model, several of these methods are analyzed in previous studies. Thus, this paper evaluates the methods based on the decision tree technique (DT, RF, LightGBM, Adaboost and XGBoost), based on the PIMA Diabetes Indian data (PID). The aim is to show the predictive ability of the methods of this technique and to determine the appropriate method for predicting diabetes with raw data. The PIMA data are described statistically, and the comparative analysis of the models is performed following K-fold cross-validation, before and after class balancing. At the end of the experiment, the best results are obtained by LightGBM, XGBoostand RF on different metrics.

Introduction:-

Diabetes is one of the most common chronic diseases in the world today, with 537 million diabetics in 2021 [1]according to the International Diabetes Federation (IDF). This number is expected to increase by 12.2% to 783 million diabetics in 2045. Defined by the WHO as a disease that prevents the body from properly using the energy provided by the food an individual consumes [2, 3]. Diabetes is now a leading cause of death if not detected and treated early. The onset of the disease is due to a lack or resistance of the hormone insulin, which is naturally produced by the pancreas and responsible for the regulated use of glucose by the body. This is because the food eaten and digested produces sugar (glucose) in the body, which is the source of energy for the functioning of various organs such as the brain, red blood cellsand others [4, 5]. This blood-borne sugar only produces energy for the

Corresponding Author:- Silue Kolo
Address:- LASTIC, ESATIC, Abidjan, Côte d'Ivoire.

organs in the presence of insulin, which is considered a key to open the entrances to the organs. So, the accumulation of sugar, not systematically regulated in the blood is since insulin does not play its role properly. Therefore, a person with this symptom is said to be diabetic.

Through modelling using mathematical approaches, research has led to an understanding of the physiological system of diabetes, which is based on the dynamics of glucose and insulin [6].

The Bergman model resulting from their study has inspired many other such studies. The authors of [7] For example, the authors of Bergman considered the natural glucose disturbing factors of eating and physical activity in diabetic scenarios (obese and non-obese). Their conclusion is that obesity is a syndrome of insulin resistance, because insulin variations in obese diabetics are disturbed and the peaks are higher.

For several years, artificial intelligence has been used to detect and predict complex diseases such as diabetes, based on existing data. In general, the effectiveness and efficiency of an artificial intelligence solution depends on the nature and characteristics of the data on the one hand and the performance of the intelligent method on the other. Machine learning and deep learning algorithms (neural network techniques) [8] are commonly selected to efficiently build data-driven intelligent systems. These algorithms are chosen according to whether they cananalyze classification or regression [9, 10]. Determining an appropriate prediction model for the target application in the fieldof medicine remains a real challenge. To face this challenge, comparative studies are conducted on several different methods, often of different techniques. However, in the literature, the selection of prediction methods is done empirically. For our study, we propose to examine the detection and prediction of diabetes using different algorithms selected by technical similarity. These are the algorithms based on the decision tree technique, namely, Decision Tree (DT), Random Forest (RF), LightGBM, AdaBoost and XGBoost, for the classification of diabetic and non-diabetic individuals.

Various works using supervised learning algorithms to classify, detect and predict diabetes from data have already been conducted on PIMA data. In many studies, decision tree methods outperform other techniques as best estimators. The authors [11] found that decision trees estimate better than neural networks, with an accuracy of 83%. The same is true for the study [12] which found that Random Forest (RF) with 98.48% accuracy, was voted the best estimator among the following methods: Support Vector Machines, K-Nearest Neighbor, Naïve Bayes, Gradient boosting and Logistic Regression. The best accuracy in the study of [5] is 79.42% by Adaboost and in the study of [13]. The best accuracy in the study of is 77.54% achieved by XGBoost. The same "decision tree" technique was used in the search for the best predictive model for diabetes [14, 15], but on data from preprocessed PIMA Indians. The authors of [16]used only LightGBM to predict diabetes from PIMA, with 95.20% accuracy of estimation. The problem generally addressed is almost the same, namely, to predict with the best accuracy to allow doctors to make a good diagnosis early, for a quick and perfect management of the diabetic. This work analyses the predictive capacity of decision trees in the sense of some studies where the methods of the tree technique were the best compared to the methods of other techniques.

In the literature, the choice of methods is not based on any criteria mentioned, except that they are supervised learning methods. Indeed, the same study can be conducted using methods based on different classification techniques. Thus, models from different techniques are compared to determine the best one. The methods used to model prediction are in the following forms: decision tree, classification rules, mathematical formulae, neural networks, naive Bayesian technique, support vector machines or K-nearest neighbors. If the tree-based methods are used very often and manage to outperform the others, then one is tempted to conduct a study involving them exclusively. This paper proposes a predictive analysis of diabetes using decision tree methods. These are those generally used in the literature for the prediction of diabetes disease, namely: XGBoost, LightGBM, etc. This work differs from previous research in the following ways. First, a single machine learning technique through five of its commonly used methods in prediction, which are evaluated using five metrics. Second, using non-preprocessed data from the PIMA Indians. The objective is to determine which of these methods produces a powerful model capable of detecting and predicting diabetes on raw, non-preprocessed data. Thus, this study could reveal the robustness of one of these algorithms in the context of using unprocessed data.

The rest of the article is organized as follows. Section 2 presents the materials and the proposed methodology. Section 3 presents the experimental results and discussion. Finally, Section 4 concludes the article, presents the limitations of the work and future work.

Materials And Methods:-

According to Han et al.[9] classification is a process of finding a model that describes and distinguishes classes of data or concepts. Classification may need to be preceded by a relevance analysis to try to identify the attributes that are significantly relevant to the classification process, as in our case. Our motivation is to propose a rigorous approach to modelling diabetes prediction on the PIMA Indian women dataset, which will be applicable to other such datasets. These data are analyzedby identifying and critically evaluating information about the disease. But our approach is to evaluate the estimators, without any preprocessing of the data, to identify the best one in terms of performance. Such an approach will be easily applicable to other data sets in the same context. The experimental work was carried out using a computer with the following characteristics Intel(R) Core (TM) i7-4700MQ CPU @ 2.40GHz 2.39 GHz, using the free Python library Scikit-learn which is intended for machine learning. The operating system is Windows 10 Professional version 21H2 of 64-bit type.

Presentation of the dataset

The PIMA Indian Diabetes dataset downloadable from Kaggle or UCI Machine Learning Repository. It is a dataset of 768 records describing Indian PIMA women (Table 1) on eight (8) explainable variables and a target variable with values of 1 (corresponding to a diabetic) and 0 (corresponding to a non-diabetic). These are quantitative, unbalanced data (Table 2) on the output variable, which represents 500 negative instances (65.1%) and 268 positive instances (34.9%).

Table 1:- Description of the PIMA Indian women dataset.

N°	Attribute name	Attribute description	Value area	Mean ± SD	Distinct values
1	Pregnancies	Number of times a woman has become pregnant	0 - 17	3,8 ± 3,3	17
2	Glucose (mg/dl)	Glucose concentration in the oral glucose tolerance test for 120 min	0 - 199	120,8 ± 31,9	136
3	BloodPressure(mmHg)	Diastolic blood pressure (when blood flows through the arteries between the heart)	0 - 122	69,1 ± 19,3	47
4	SkinThickness(mm)	Triceps skinfold thickness (mm), determined by collagen content	0 - 99	20,5 ± 15,9	51
5	Insulin (mu U/mL)	Serum insulin for 2 hours	0 - 846	79,7 ± 115,2	186
6	BMI (kg/m2)	Body mass index (weight/(height)^2)	0 - 67,1	31,9 ± 7,8	248
7	DiabetesPedigreeFunction	Attractive attribute used in the prognosis of diabetes Function	0,078 - 2,42	0,4 ± 0,3	517
8	Age	Number of years of age	21 - 81	33,2 ± 11,7	52

Table 2:- The unbalanced proportions of output values

Class number	Labelle	Number	Percentage
0	Non-diabetic	500	65,1%
1	Diabetic	268	34,9%.

In this dataset, non-diabetics are more frequent than diabetics **(Table 2).**The disadvantage is that the model could learn on a training dataset that is mostly composed of non-diabetics, which may influence its performance in correctly predicting a case of diabetes. This has the consequence of influencing its performance in correctly predicting a diabetic case. That is, the results of the performance measures (accuracy, recall score, precision score, AUC_ROC score and F1 measure) are likely to be influenced as a whole. To show this, performance is assessed before and after a class balancing step.

The null values (0) of the Indian PIMA, Glucose (5), BlooPressure(35), SkinThickness(227), Insulin (374) and BMI (11) dataset characteristics are identified as missing values (Table 3). But this study was conducted with data containing missing values in their original state, considered as null values.

Table 3:- Missing values (MV) in Indian PIMA data.

Features	Number of VMs	Percentage of VMs
Insulin	374	48,7%
Skin thickness	227	29,56%
Blood pressure	35	4,56%
BMI	11	1,43%
Glucose	5	0,65%

Modelling

Prediction methods:

The problem to be addressed here is the classification analysis to detect whether a woman is diabetic or non-diabetic. The problem is therefore to select methods suitable for the classification analysis technique. Classification is a supervised learning task in artificial intelligence, referring to a predictive modelling problem, where a class label is predicted for a given example [9]. Mathematically, it maps a function (f) of input variables (X) to output variables (Y) as target, label or categories. In our study, the chosen machine learning methods are exclusively based on the decision tree technique, before and after class balancing. Furthermore, we did not perform any preprocessing on the data before modelling. Thus, this allows to just appreciate the effectiveness and efficiency of the model produced by the performance of the modelling method. We present below the methods used in our study.

Decision tree

A decision tree (DT) is a widely used supervised machine learning method for dealing with classification and regression problems [17, 18]. ID3 [18] , C4.5 [19] and CART [20] are well known for DT algorithms. Its representation is an acyclic directed graph whose nodes correspond to the variables chosen based on quality criteria, while the arcs represent the modalities of a predictor variable. The aim of implementing a decision tree is to separate the classes in such a way as to obtain homogeneous leaves in terms of class. Instances are classified by checking the attribute defined by that node, starting with the root node of the tree, then moving down the branch of the tree corresponding to the value of the attribute. For splitting, the most popular criteria are "Gini" (1) for Gini impurity and "entropy" (2) for information gain which can be expressed mathematically as [17].

$$\text{Entropy: } H(x) = -\sum_{i=0}^{n} p(x_i)\log_2 p(x_i) \quad (1)$$

$$\text{Gini}(E) = 1 - \sum_{i=0}^{c} {p_i}^2 \quad (2)$$

Random Forest

The random forest (RF) is an extension of a decision tree [21]. It is a method formed by the combination of several decision tree predictors, by the bagging method [22]. In this case, the decision tree prediction is based on a randomly selected subset of the decision trees. In this case, the model prediction is the class that obtains more majority votes,or the averages of the results obtained by all the trees. This minimizes the overfitting problem and increases the accuracy and control of the prediction [17]. With better input features, the random forest can achieve very high performance [23].

Adaboost

Adaboost(adaptive boosting) is a meta-algorithm for boosting [24] that relies on the iterative selection of weak classifiers based on a distribution of training examples. It is the method that allows poor classifiers to improve by learning from their mistakes, thus creating a powerful, high accuracy classifier [25]. Each example is weighted according to its difficulty with the current classifier. While the random drill (RF) uses parallel assembly, Adaboost uses sequential assembly. Adaboost is used to boost the performance of decision trees on binary classification problems. However, it can be sensitive to noisy data [17].

XGBoost

XGBoost(eXtreme Gradient Boosting) implements the gradient boosting algorithm for decision trees. It is a gradient boosting algorithm that is composed of "gradient" and "boost", like the Randomized Forest (RF) [21]. The gradient consists of minimizing the loss function, in the same way as neural networks. Boosting combines weak classifiers (processes that make inaccurate serial judgements) which are here decision trees. XGBoost is one of the ensembles learning algorithms which involves using multiple decision trees to build the prediction [17]. It is an algorithm that improves the accuracy of predictions by using the trees in a specific order. For an observation, each tree gives a result, and the final prediction is obtained by summing each of the obtained values given by the trees.

LightGBM

LightBGM(Light Gradient Boosting Machine) implements a decision tree-based gradient boosting algorithm to increase model efficiency and reduce memory usage. The LightGBM algorithm is characterized by the two techniques of GOSS: Gradient-based One-Side Sampling and EFB: Exclusive Feature Bundling. LightGBM is called "Light" because of its computing power and faster results. In order to reduce implementation time, a team at Microsoft developed the Light Gradient Boosting Machine (LightGBM) in April 2017 [26]. The main difference is that the decision trees in LightGBM are grown per leaf, instead of checking all previous leaves for each new leaf.

k-fold cross-validation

The cross-validation method (K-fold with k=5, 7, 12) was used to validate the performance of the model [10]. The choice of the three k-fold values was made after an evaluation procedure of the machine learning methods on the set of metrics for k varying between 3 and 14. It appears that with k=5, 7, 12, the estimators gave the best performance measure values.

The model's measurements:

Equations 3, 4, 5, 6 and 7 below were used to calculate model accuracy, precision score, recall, F1 function and area under the ROC curve (AUC) respectively. Each evaluation metric quantifies the performance of the predictive model [27], [9]:

- Accuracy is defined as the ratio of well ranked examples to the total number of examples. It is the performance of the model in predicting on the unobserved data (correct prediction rate). It is the metric that will globally quantify the model, measuring its ability to correctly predict both diabetics and non-diabetics.

$$\text{accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

- Accuracy is the proportion of items that are well classified for a given class. It is the ability of the classifier not to label a non-diabetic as diabetic.

$$\text{precision} = \frac{TP}{TP+FP} \quad (4)$$

- Sensitivity is the conditional probability of having a positive test result if the individual has diabetes. It represents the frequency of positive test responses. For example, how many of those predicted to be diabetic are? It therefore measures the capacity of a model to detect patients. The closer the sensitivity is to unity, the fewer the errors in detecting sick subjects (false negatives).

$$\text{Recall(sensitivity)} = \frac{TP}{TP+FN} \quad (5)$$

- The f1 function is the overall measure of the precision of a model that combines precision and recall. It is the weighted harmonious average of precision and recall. A good F1 means that there are low false positives and low false negatives. An F1 score is considered perfect when it is 1, while the model is a total failure when it is 0.

$$F1 = \frac{2*(\text{precision} * \text{rappel})}{(\text{precision} + \text{rappel})} \quad (6)$$

where

- true positive (TP) represents the number of diabetics identified in the corresponding set of diabetics,
- true negative (TN) means the number of non-diabetics classified in the non-diabetic set,
- The false positive (FP) is the number of diabetics identified in the set of non-diabetics,
- and false negatives (FN) represent the number of non-diabetics identified in the diabetic set.

- The CUA is one of the popular ranking-type measures. In [28]the CUA was used to build an optimized learning model and to compare learning algorithms [29]. Unlike the threshold and probability metrics, the AUC value reflects the overall ranking performance of a classifier. For a two-class problem [28]problem, the AUC value can be calculated as follows

$$AUC = \frac{n_p - \frac{n_p(n_n+1)}{2}}{n_p n_n} \quad (7)$$

where S_p is the sum of all positive ranked examples, while n_p and n_n are the number of positive and negative examples respectively. The AUC has been shown to be theoretically and empirically better than the accuracy metric [30] for evaluating the performance of the classifier and discriminating an optimal solution during classification training.

Results And Discussion:-

Descriptive analysis of the data

The average distribution of the input data by class (Table 4) allows the analysis of the average value that expresses each class. All variables contain quantitative data. Furthermore, the nature of the data per variable and their distribution with respect to the target variable are summarized (Table 5) by different empirical indicators of central tendency or dispersion.

Table 4:- Average distribution of input data by variable for each class.

Class	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
0	3,30	110,64	70,88	27,24	130,29	30,86	0,43	31,19
1	4,87	142,32	75,32	33,00	206,85	35,41	0,55	37,07

Table 5:- Statistical summary of the PIMA Indian women's diabetes data set.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768	768	768	768	768	768	768	768	768
mean	3,85	120,89	69,11	20,54	79,80	31,99	0,47	33,24	0,35
std	3,37	31,97	19,36	15,95	115,24	7,88	0,33	11,76	0,48
min	0	0	0	0	0	0	0,08	21	0
25%	1	99	62	0	0	27,30	0,24	24	0
50%	3	117	72	23	30,50	32	0,37	29	0
75%	6	140,25	80,00	32,00	127,25	36,60	0,63	41	1
max	17	199	122	99	846	67,1	2,42	81	1
CV	87,63	26,45	28,01	77,68	144,42	24,64	70,22	35,38	136,68

Table 6:- Correlation coefficient between input and output variables.

Features	**Correlation coefficient**
Outcome	1,00
Glucose	0,47
BMI	0,29
Age	0,24
Pregnancies	0,22
DiabetesPedigreeFunction	0,17
Insulin	0,13
SkinThickness	0,07
BloodPressure	0,07

Table 6shows that all but two of the input variables (SkinThickness and BloodPressure) have a non-negligible correlation with the target variable. This demonstrates that these variables can provide knowledge about the disease of diabetes.

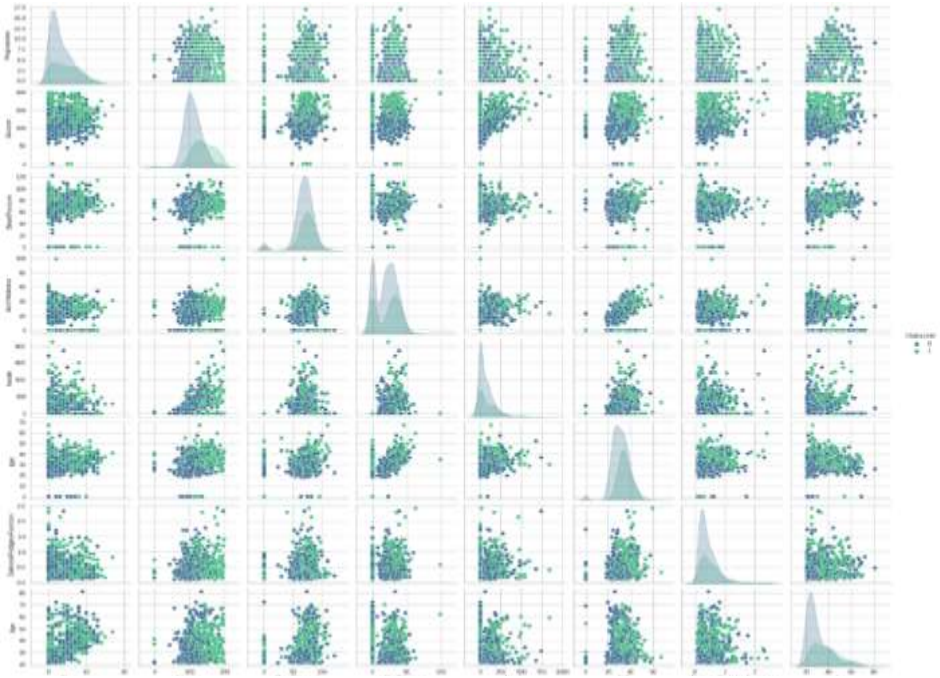

Figure 1:- Feature Scatterplot Matrix.

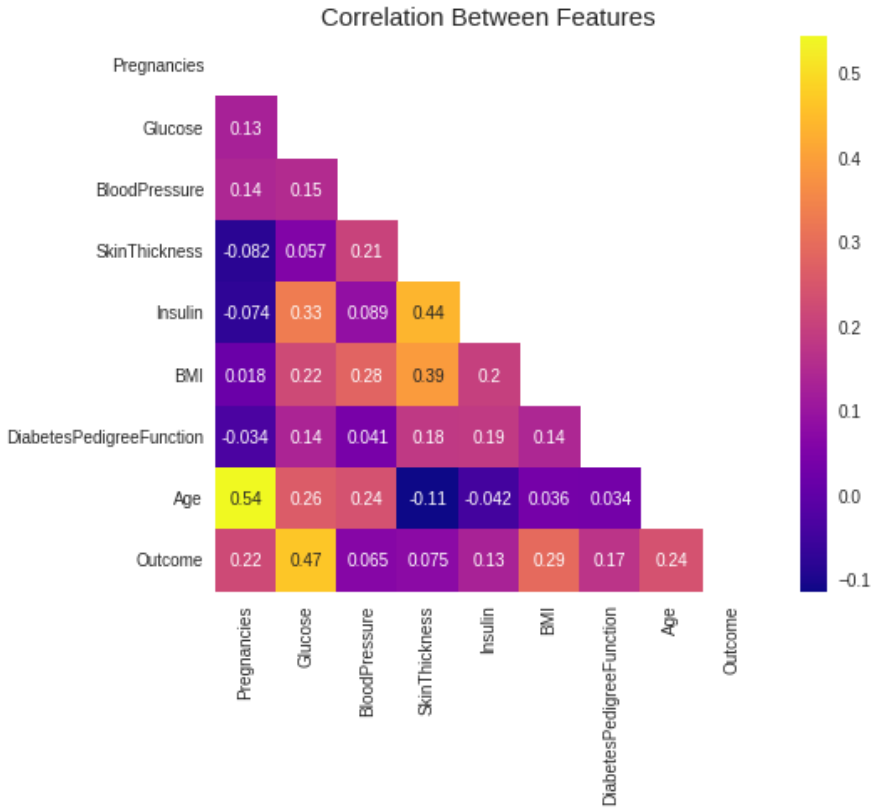

Figure 2:- Correlation matrix between variables.

Table 2shows a balancing problem between the data and Table 3 shows the significant presence of missing values, here represented by zero (0). In summary, a database quality problem clearly emerges. In machine learning, all this can influence the predictive performance of the model.

The scatterplot matrix is useful for identifying the level of dispersion of the data by pair of entities in a preliminary way. If the points are scattered, it means that there is no obvious relationship, whereas if the points are roughly in a straight line, it shows that they are linearly related. Figure 1, shows that the most closely correlated characteristics include [pregnancy and age], [glucose and age], [skin thickness and BMI], [blood pressure and MHI], [skin thickness and insulin] and [glucose and insulin]. Thus, we note a positive correlation by the numbers (Figure 2). On the diagonal, the distribution of classes for all attributes is shown (Figure 1), which is not linearly separable. All this shows that the prediction of diabetes is a difficult task. Figure 3 shows the density and frequency of the data distribution per variable.

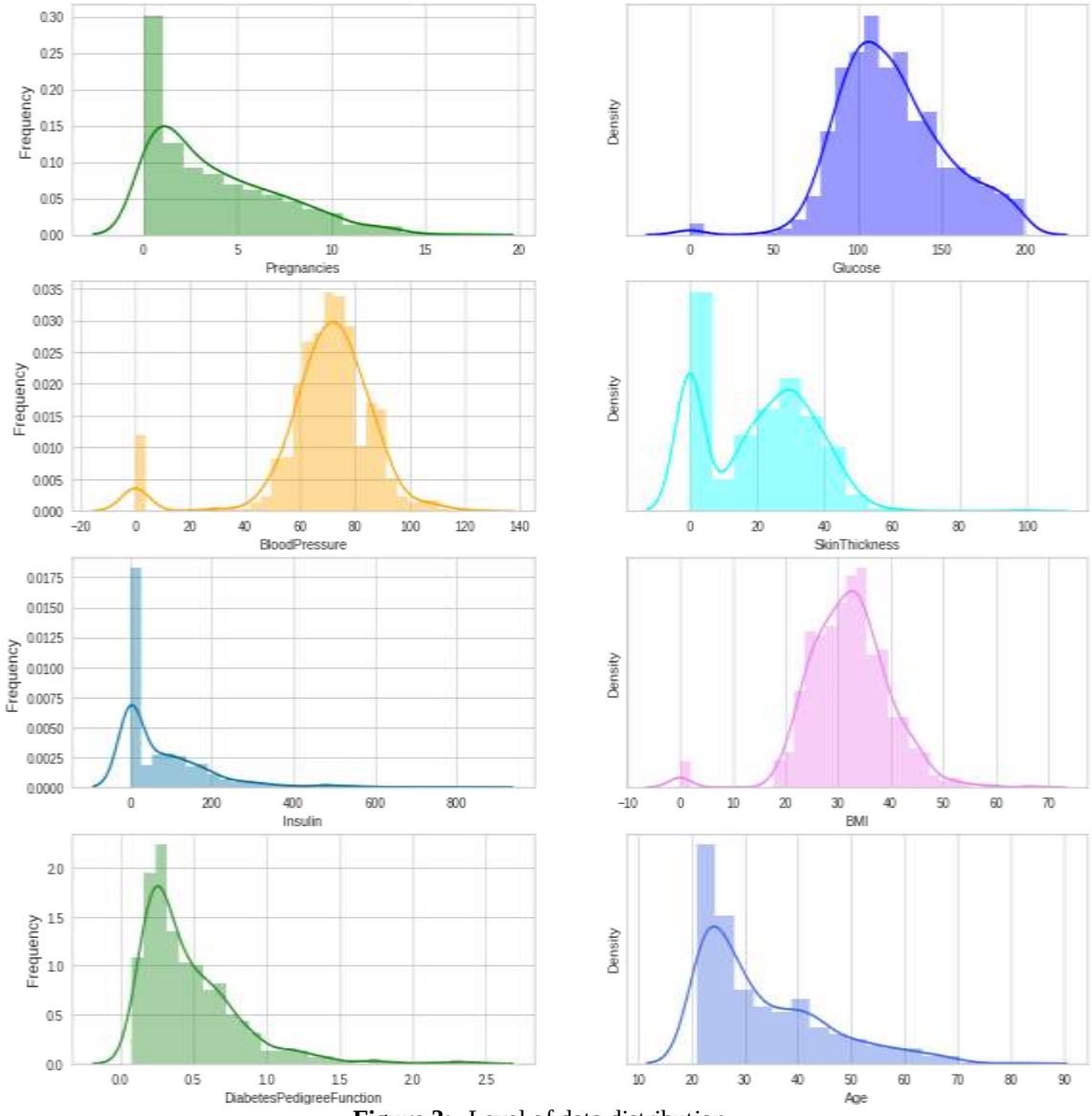

Figure 3:- Level of data distribution.

Predictive analysis of PIMA data

Many studies have evaluated the performance of a machine learning model using the accuracy that indicates the cases correctly predicted by the model. In our case, it is rather preferable to measure the number of positive cases that the model can correctly predict to stop the spread of diabetes. The reliability of our model should be verified by the number of positive cases predicted, thus by the measures of precision and recall. Recall is a useful measure in cases where false negatives are of greater concern than false positives. It does not matter if a false alarm is triggered in our study. But real positive cases should not go unnoticed. The AUC is used to discriminate the optimal and overall performing solution.

K-fold (k=5)

Table 7:- Comparison between the classification results of the methods, before balancing.

	f1	**recall**	**precision**	**roc_auc**	**accuracy**
Xgboost	0,6380	0,6270	0,6537	0,7235	0,7527
Decision Tree	0,5899	0,5747	0,5953	0,6844	0,7045
LightGBM	**0,6616**	**0,7016**	0,6290	**0,7388**	0,7501
RandomForest	0,6371	0,5633	**0,7195**	0,7195	**0,7683**
AdaBoost	0,6371	0,5934	0,6942	0,7237	0,7631

Table 8:- Comparison between the classification results of the methods, after SMOTE.

	f1_score	**recall_score**	**precision_score**	**roc_auc_score**	**accuracy_score**
Xgboost	**0,823712**	**0,862**	0,789815	0,817	0,817
Decision Tree	0,746908	0,754	0,740859	0,746	0,746
lightGBM	0,780381	0,794	0,770010	0,777	0,777
RandomForest	0,827620	0,858	**0,800380**	**0,822**	**0,822**
AdaBoost	0,769826	0,768	0,773474	0,771	0,771

For the K-fold cross-validation with K=5, before class balancing (Table 7), LightGBM is the best model for having obtained the best values of AUC (0.7388), recall (0.7016) and F1 measure (0.6616), even if this model is less accurate than RF. But after balancing, RF still has the best precision and AUC, followed by the XGBoost model (Table 8).

K-fold (k=7)

Table 9:- Comparison between the classification results of the methods, before balancing.

	f1_score	**recall_score**	**precision_score**	**roc_auc_score**	**accuracy_score**
Xgboost	**0,652019**	0,637941	0,670869	**0,732982**	0,761790
Decision Tree	0,587471	0,582418	0,580209	0,666933	0,701942
lightGBM	0,651644	**0,686331**	0,622162	0,729232	0,742226
RandomForest	0,647952	0,604203	**0,722866**	0,731487	**0,768319**
AdaBoost	0,610999	0,570561	0,662814	0,706265	0,747468

Table 10:- Comparison between the classification results of the methods, after SMOTE.

	f1_score	**recall_score**	**precision_score**	**roc_auc_score**	**accuracy_score**
Xgboost	0,810596	**0,844316**	0,781630	0,804033	0,804069
Decision Tree	0,743985	0,752319	0,740114	0,742036	0,742054
lightGBM	0,787001	0,812179	0,765026	0,779916	0,780016
RandomForest	**0,811897**	0,844260	**0,783529**	**0,804983**	**0,805019**
AdaBoost	0,750422	0,744299	0,759495	0,753004	0,753015

For K-fold cross-validation with K=7, before class balancing (Table 9), LightGBMshows the best recall (0.6863), while the best AUC (0.7329) and the best F1 measure are obtained by XGBoost. The best accuracy is still achieved by RF. But after balancing (Table 10), the best model is RF with the best values for precision, AUC and F1 measurement, followed by XGBoost.

K-fold (k=12)

Table 11:- Comparison between the classification results of the methods, before balancing.

	f1_score	**recall_score**	**precision_score**	**roc_auc_score**	**accuracy_score**
Xgboost	0,644516	0,630105	0,671693	0,728888	0,759115
Decision Tree	0,588765	0,589592	0,592628	0,684822	0,713542
lightGBM	0,636862	**0,678524**	0,606279	0,719198	0,731771
RandomForest	**0,655442**	0,619071	**0,708554**	**0,737333**	**0,773438**
AdaBoost	0,629605	0,597167	0,680248	0,717355	0,753906

Table 12:- Comparison between the classification results of the methods, after SMOTE.

	f1_score	**recall_score**	**precision_score**	**roc_auc_score**	**accuracy_score**
Xgboost	0,814156	**0,842625**	0,793237	0,810371	0,810253

Decision Tree	0,782537	0,798490	0,771730	0,780294	0,780180
lightGBM	0,770427	0,786779	0,761490	0,768438	0,768192
RandomForest	**0,825563**	0,842383	**0,815062**	**0,824235**	**0,824166**
AdaBoost	0,776617	0,768438	0,790997	0,778286	0,778184

For K-fold cross-validation with K=12, both before (Table 11) and after class balancing (Table 12), the best model remains RF followed by XGBoost.

From the different stages of experimentation conducted in this study, we note that the algorithms perform better in the case of class balancing regardless of the K-fold value for cross-validation. As the classes were balanced by the SMOTE [31]technique, this not only improved the performance of all the algorithms studied, but also allowed us to determine the best model with RF. Indeed, RF obtains the best measures for accuracy (82.41%), precision (81.50%), F1 measure (0.8255) and AUC (0.8242) and the second-best recall (0.8423), after XGBoost. Furthermore, all algorithms can be good estimators of diabetes from the PIMA dataset, despite using raw data. A method based on the decision tree technique would be moderately sensitive to missing values. The best measurements are nevertheless obtained with K-fold cross-validation (K=12) by RF. This confirms the results obtained by [32] which already indicated that tree algorithms have been widely used in several disciplines because they are easy to use, unambiguous and robust even in the presence of missing values. This may explain the fact that these algorithms outperform others in comparative studies. The results of better estimators that are DT in [11]RF in [12], Adaboost in [5], LightGBM in [16] and XGBoost in [13]give us comfort in the choice of tree methods to conduct our study. As the aim of the study is to detect almost all diabetics and minimize the false negative rate, we prefer the classifier that achieves good sensitivity. This is either RF, XGBoost or LightGBM depending on the case of class balancing or not, for the value of K=5, 7, 12 of the cross validation. Since the model performs well when the area under the curve is high, it is still between these three algorithms that the best model is determined depending on the case. But for us, the appropriate model is obtained by RF, which performs best on all measurements (Table 12) with cross-validation (k=12).

Conclusion:-

In the management of diabetes, successfully detecting and predicting it to curb its rise in the world has been a research problem for many years. Machine learning and data mining are technologies of great importance, which are widely used to model such a complex task. For example, an intelligent detection system based on available data can be used to help doctors diagnose diabetes. This will allow many people to avoid getting diabetes in their lifetime. But the aim is to obtain an effective and efficient model through methods based on the decision tree technique. In the context of comparing the performance of the methods, and in order to explain and interpret the model obtained, we have used the methods that have a technical similarity in prediction. Our study shows that among the classifiers based on the tree technique, without data preprocessing, LightGBM, XGBoostand Random Forest are the best. We recommend their use in similar studies. This result was obtained on the PIMA Indian dataset without the preprocessing of the data, which does have missing values. But, by balancing the classes under these conditions, depending on whether K=5, 7, 12, the performance of the different methods improved. At the K-fold validation (12), the results are better, and the appropriate model is obtained with RF. Nevertheless, this remains to be confirmed in another study, this time with the preprocessing of the data on the same PIMA Indian dataset.

References:-

[1] International Diabetes Federation, "IDF_Atlas_10th_Edition_2021".
[2] WHO, "Definition, Diagnosis and Classification of Diabetes mellitus and its Complications," NCD/NCS/99.2, 1999.
[3] T. Mathie, B. Amélie, F. Philippe, and A. Amar, "Pathophysiology of diabetes," 2018.
[4] L. Bellamy, J. P. Casas, A. D. Hingorani, and D. Williams, "Type 2 diabetes mellitus after gestational diabetes: a systematic review and meta-analysis," The Lancet, vol. 373, no. 9677, pp. 1773-1779, 2009, doi: 10.1016/S0140-6736(09)60731-5.
[5] J. J. Khanam and S. Y. Foo, "A comparison of machine learning algorithms for diabetes prediction," ICT Express, vol. 7, no. 4, pp. 432-439, Dec. 2021, doi: 10.1016/j.icte.2021.02.004.
[6] R. N. Bergman, "Toward Physiological Understanding of Glucose Tolerance Minimal-Model Approach," 1989, Accessed: Nov. 08, 2022. [Online]. Available: http://diabetesjournals.org/diabetes/article-pdf/38/12/1512/356697/38-12-1512.pdf

[7] K. Silue, H. K. Konan, M. Coulibaly, and O. Asseu, "Determination of a Numerical Analysis Algorithm for the Regulation of Blood Sugar in Diabetics," Open Journal of Applied Sciences, vol. 11, no. 08, pp. 908-928, 2021, doi: 10.4236/ojapps.2021.118067.
[8] I. H. Sarker, "Data Science and Analytics: An Overview from Data-Driven Smart Computing, Decision-Making and Applications Perspective," SN Computer Science, vol. 2, no. 5. Springer, Sep. 01, 2021. doi: 10.1007/s42979-021-00765-8.
[9] Jiawei Han, Micheline Kamber, and Jian Pei, "Data Mining Third Edition," 2012.
[10] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, "Credibility: Evaluating what's been learned. Data mining: Practical machine learning tools and techniques, 143-186," 2005.
[11] E. PekelÖzmen and T. Özcan, "Diagnosis of diabetes mellitus using artificial neural network and classification and regression tree optimized with genetic algorithm," J Forecast, vol. 39, no. 4, pp. 661-670, Jul. 2020, doi: 10.1002/for.2652.
[12] D. Jashwanth Reddy et al , "Predictive machine learning model for early detection and analysis of diabetes," Mater Today Proc, Oct. 2020, doi: 10.1016/j.matpr.2020.09.522.
[13] I. Gnanadass, "Prediction of Gestational Diabetes by Machine Learning Algorithms," IEEE Potentials, vol. 39, no. 6, pp. 32-37, Nov. 2020, doi: 10.1109/MPOT.2020.3015190.
[14] S. Habibi, M. Ahmadi, and S. Alizadeh, "Type 2 Diabetes Mellitus Screening and Risk Factors Using Decision Tree: Results of Data Mining," Glob J Health Sci, vol. 7, no. 5, Mar. 2015, doi: 10.5539/gjhs.v7n5p304.
[15] A. A. al Jarullah, "Decision tree discovery for the diagnosis of type II diabetes," in 2011 International Conference on Innovations in Information Technology, Apr. 2011, pp. 303-307. doi: 10.1109/INNOVATIONS.2011.5893838.
[16] B. Shamreen Ahamed and M. Sumeet Arya, "LGBM Classifier based Technique for Predicting Type-2 Diabetes."
[17] Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, and Bertrand Thirion, "Scikit-learn_Machine Learning in Python," Journal of Machine Learning Research 12, 2011.
[18] j.r. Quinlan, "Introduction of Decision Trees," 1986.
[19] J. Ross Quinlan, "C4.5 Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc. 1993 - Programs for Machine Learning," 1993.
[20] Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone, "Classification and Regression Tree," 1984.
[21] LEO BREIMAN, "Random Forests," 2001.
[22] LEO BREIMAN, "Bagging Predictors," 1996.
[23] I. H. Sarker, A. S. M. Kayes, and P. Watters, "Effectiveness analysis of machine learning classification models for predicting personalized context-aware smartphone usage," J Big Data, vol. 6, no. 1, p. 57, Dec. 2019, doi: 10.1186/s40537-019-0219-y.
[24] Yoav Freund and Robert E. Schapire, "Experiments with a New Boosting Algorithm," 1996.
[25] Arif-Ul-Islam and S. H. Ripon, "Rule Induction and Prediction of Chronic Kidney Disease Using Boosting Classifiers, Ant-Miner and J48 Decision Tree," in 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), Feb. 2019, pp. 1-6. doi: 10.1109/ECACE.2019.8679388.
[26] GuolinKeet al , "Light_2A Hightly Efficient gradient Boosting Decision Tree," 2017.
[27] Witten IH, Frank E, Trigg LE, Hall MA, Holmes G, and Cunningham SJ, "Weka _ Practical Machine Learning Tools and Techniques with Java Implementations," 1999.
[28] David J. Hand and Robert J. Till, "A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems," 2001, doi: 10.1023/A:1010920819831.
[29] F. Ridzuan and W. M. N. Wan Zainon, "Diagnostic analysis for outlier detection in big data analytics," Procedia Comput Sci, vol. 197, pp. 685-692, 2022, doi: 10.1016/j.procs.2021.12.189.
[30] Jin Huang and C. X. Ling, "Using AUC and accuracy in evaluating learning algorithms," IEEE Trans Knowl Data Eng, vol. 17, no. 3, pp. 299-310, Mar. 2005, doi: 10.1109/TKDE.2005.50.
[31] V. , C. Nitesh, W. , B. Kevin, O. , H. Lawrence, and P. K. W., "SMOTE (Synthetic Minority Over-sampling Technique)," 2002.
[32] Trevor Hastie, Robert Tibshirani, and Jerome Friedman, "The Elements of Statistical Learning: data mining, inference, and prediction," 2009.