



# Using Machine learning, several types of model-based type 2 Diabetes Mellitus prediction

By

**Rahima Akter Runa**

**171-35-230**

**Submission Date:** 05 /06/21

**Department of Software Engineering  
DAFFODIL INTERNATIONAL UNIVERSITY**

## BOARD OF EXAMINERS

SIGNATURE

Name

**Chairman**

Title

Department of Software Engineering

Daffodil International University

SIGNATURE

Name

**Internal Examiner**

Title

Department of Software Engineering

Daffodil International University

SIGNATURE

Name

**Internal Examiner**

Title

Department of Software Engineering

Daffodil International University

SIGNATURE

Name

**External Examiner**

Title

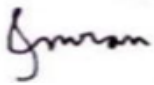
Department name

Institution

## THESIS DECLARATION

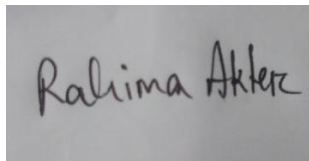
This report was written by me, Rahima Akter Runa, under the supervision of Md. Rajib Mia, Lecturer, Dept. of Software Engineering in Daffodil International University. I further attest that no part of this study, or any part of it, has been submitted for a degree anywhere else.

**Supervised by,**



Dr. Imran Mahmud  
Associate Professor & Head  
Department of Software Engineering  
Faculty of Science and Information Technology  
Daffodil International University

**Submitted by,**



**Rahima Akter Runa**

**ID: 171-35-230**

Department of Software Engineering  
Faculty of Science and Information Technology  
Daffodil International University

## ACKNOWLEDGEMENT

Firstly, I am offering my thanks to god-like Allah for enabling me to work this proposal work. I might want to offer our genuine thanks to my fair boss, MD. Rajib Mia, Lecturer, Department of Software Engineering. This theory would not have been finished without his help and direction. His consistent support gave me the certainty to complete my work. I might likewise want to give exceptional appreciation to Bikash Kumar Paul, Lecturer at Information and Communication Technology in Mawlana Bhashani Science and Technology University and Shariful Islam, Lecturer, Department of Software Engineerin. Their legitimate course and direction assist me with setting up this proposal work with no trouble. I offer my heartiest thanks towards the whole division of Software Engineering at Daffodil International University for giving well-rounded schooling and information. I additionally offer my thanks to all my instructor's SAM Matiur Rahman, Associate Professor; Dr. Imran Mahmud, Professor, and Head, Dept. of Software Engineering. I likewise need to thank our decent MD. Sk Fazlee Rabby sir, Nayeem Ahmed Sir, and Raihana Zannat mam. They were consistently sincere to us, helped us, and persuaded us. We are consistently appreciative to them The information that I have gained from the classes in my level of a single guy's in computer programming level was fundamental of this proposition. In the course of directing the examination, the important data was gathered through books, diaries, electronic media, and other optional sources. I additionally need to thank my companions for giving me backing and consolation. Their hopefulness and support have permitted them to defeat any obstruction at any stage.

## TABLE OF CONTENT

<b>Contents</b>	<b>Page</b>
THESIS DECLARATION .....	i
ACKNOWLEDGEMENT .....	ii
TABLE OF CONTENT .....	iii
LIST OF TABLES.....	iv
LIST OF FIGURES .....	iv
LIST OF ABBREVIATION.....	v
ABSTRACT .....	vi
CHAPTER 1: INTRODUCTION .....	1
1.1 Background: .....	1
1.2 The Research Project's Motivation: .....	2
1.3 Problem Statement: .....	2
1.4 Research Question: .....	3
1.5 Research Objectives: .....	3
1.6 Research Scope: .....	3
1.7 Thesis Organization: .....	4
CHAPTER 2: RELATED WORK REVIEW .....	5
2.1 Related work: .....	5
CHAPTER 3: Describe Machine Learning Algorithm in a few word.....	5
3.1 KNN: .....	5
3.2 Logistic Regression: .....	6
3.3 Linear Regression: .....	6
3.4 Decision Tree Classifier: .....	6
3.5 Gradient Boosting Classifier: .....	6
3.6 Light Gradient Boosting Machine: .....	6
CHAPTER 4: Methodologies.....	8
4.1 Data Description: .....	8
4.2 Preprocessing of dataset: .....	8
4.2.1 Data Analysis: .....	8
4.2.1 Data Imputation: .....	8

4.2.1 Design and Implementation: .....	8
CHAPTER 5 Result & Discussion: .....	16
CHAPTER 6 Conclusion: .....	16
5.1 Findings: .....	16
5.2 Recommendation for Future works: .....	16
REFERENCES .....	16

## LIST OF TABLES

Table 01	Dataset describe	11
Table 02	Different ways hold a confusion matrix	
Table 03	Qualities for various measures for various classification techniques	

## LIST OF FIGURES

Figure 01	Decision Tree	5
Figure 02	Dataset used in this study	8
Figure 03	Flowchart of this Research	9

## LIST OF ABBREVIATION

DM = Diabetes Mellitus

ECG = Electrocardiogram

PIDD = Pima Indian Diabetes Dataset

KNN = K-Nearest Neighbor

TP = True Positive

TN = True Negative

FN = False Positive

FP = False Negative

## **Abstract**

A medical services framework utilizing current registering strategies is the most elevated investigated region in medical services research. Specialists in the field of processing and medical care are steadily cooperating to prepare such frameworks for more innovation. Diabetes is considered as one of the deadliest and ongoing sicknesses it prompts inconveniences like visual impairment, removal, and cardiovascular infections in a few nations, and every one of them is attempting to forestall this illness at the beginning phase by diagnosing and anticipating the indications of diabetes utilizing a few strategies. The thought process of this examination is to look at the exhibition of some Machine Learning calculations, used to anticipate type 2 diabetes infections. In this paper, we apply and assess six Machine Learning calculations (Logistic Regression, Decision Tree, Linear Regression, K-Nearest Neighbors, Light Gradient Boosting Machine, and Gradient Boosting Machine) to foresee patients with or without type 2 diabetes mellitus. These procedures have been trained and tested on a notable Pima Indian dataset. The exhibitions of the tested calculations have been assessed for this situation dataset with boisterous information (before pre-processing/some information with missing values) and dataset set without boisterous information (after pre-processing). The outcomes analyzed utilizing distinctive similitude measurements like Accuracy, Sensitivity, and Specificity give the best presentation with worry to best in class.

# Chapter 1

## Introduction

### 1.1 Background

Diabetes is a term used to describe a set of metabolic disorders problems marked by elevated Glucose levels in the blood over a lengthy length of time. Diabetes Mellitus is another name for the condition. High glucose symptoms include frequent micturition, a dry sensation, and an increased desire to eat [1]. Diabetes, if not treated immediately, death might occur, diabetic ketoacidosis, and hyperosmolar hyperglycemic state. It may resulting in long-term complications such as cardiovascular infection, cerebrum stroke, renal disappointment, foot ulcers, and eye strain, among others [2]. Diabetes develops when the body's pancreas malfunctions does not produce enough insulin as well as when the insulin that is supplied to the cells and tissue of the body is not receive. There are 3 forms of DM:

- Type 1 diabetes can strike at any age, although it strikes more frequently in children and teenagers. Type 1 diabetes is a severe condition that is resistant to treatment. People with type 1 diabetes require an exogenous insulin supplement to compensate for the pancreas' reduced insulin production.
- In adults, it is more common to have type 2 diabetes, accounting for over 90% of all diabetes occurrences. The most prevalent symptom is insulin blockage, which occurs when the body does not properly respond to insulin. This can eventually deplete the pancreas, leading the body to produce less and less insulin, resulting in very higher glucose levels in certain people with type 2 diabetes.
- Gestational diabetes discovered in a pregnant woman who did not have diabetes before becoming pregnant. A few women have had gestational diabetes affect more than one pregnancy [3].

If they have a family history of type 2 diabetes, their chances of developing it are higher:

- Anyone with a BMI more than 25, regardless of age, who has additional risk factors such as hypertension, abnormal cholesterol levels, an inactive lifestyle, a family history of polycystic ovary disease or cardiovascular disease, and who has a direct link to diabetes [4].
- Anyone over the age of 45 is urged to have a basic glucose test and, if the results are normal, to be checked on a regular basis after that.



- Women who have prediabetes are urged to be tested for diabetes on a regular basis.
- Anyone who's been diagnosed with prediabetes is urged to get tested every year.

When a specialist determines that a person has prediabetes, they advise them to improve their lifestyle. Accepting a health framework and a balanced eating regimen will help to avoid diabetes [5].

This examination plans to decide the danger of the improvement of diabetes in a person. The resulting segments of the article comprise related work in segment 2. In segment 3, a brief description of the machine learning algorithms used is provided. The technique is depicted in area 4 while brings about segment 5. Area 6 sums up the end.

## 1.2 Motivation of Research

Type 2 diabetes is caused by two issues that are intertwined: Insulin resistance develops in tissue, fat, and vital organs. A few works, I showed, are centered on enhancing K-means by developing the bunch community's established system. But, unlike the logistic regression, my revised model is based on the basis for predicting DM2. Despite the fact that the modified model isn't quite as confusing as the original, it still achieves impressive results. The main difficulties I addressed were enhancing the precision of the expectation model and causing it to adapt to different datasets. In this research, I argue that my suggested model has a greater expected exactness than the exploratory results of other scientists.

## 1.3 Research Question:

Finding out how well Logistic Algorithm can predict Types 2 Diabetes Mellitus in whole Dataset.

How fast Logistic Algorithm for predicting Diabetes type 2.

Predict and Count Type 2 Diabetes positive and negative.

Why Logistic Algorithm is better than others algorithms.

## 1.4 Research Objectives:

My research goal was to develop a model that can predict the Type 2 Diabetes Mellitus from Pima Indian Diabetes Dataset, and which Model gave the best accuracy among all models.

Finally Logistic Algorithm gave the best prediction for the using data.

## 1.5 Research Scope:

### Thesis Organization:

Chapter 1: In this chapter, I discuss the introduction of our thesis. Here we also discuss our research objective and our research questions.

Chapter 2: In this chapter, I discussed literature review and previous work which are related to this work.

Chapter 3: In this chapter, I discuss research Describe Machine Learning Algorithm in a few words.

Chapter 4: In this chapter, I discuss Methodologies part.

Chapter 5: In this chapter, I discuss result and discussion.

Chapter 5: In this chapter, I represent the conclusion, and future plan

## CHAPTER 2

### 2.1 Related work

As of late, machine learning algorithm has become more popular as a tool for predicting the possibility of sickness. Scientists have produced and focused numerous computations and tool sections. The huge capability of this examination sector has been highlighted in these. A few of major works that are clearly associated with the suggested problem are discussed in this part [6].

For diabetes detection, the Pima Indians Diabetes Dataset has been used in numerous research investigations. (PIDD). AI techniques and the Weka apparatus were applied. Neural organization, Machine learning algorithm, hybrid methods, and data mining processes are some of the different methods used by professionals.

Swapna et. al. in [7] applied deep learning techniques on ECG data to reveal diabetes. And they discovered 95.4% accuracy rate.

Sisodia et al. in [8] used 3 machine learning techniques on pima dataset: naïve bayes , decision tree, and support vector machine. Naive Bayes was determined to be 76.30% accurate.

Nongyao et al. in [9] examined four different decision-making systems including logistic regression, decision trees, logistic regression, and ANN. All were given greater shooting and boosting, and random forest was also incorporated. The highest level of precision attained by completely was somewhere between 84 and 86 percent.

## CHAPTER 3

### Describe Machine Learning Algorithm in a few words

#### 3.1. KNN

KNN is one of Machine Learning's most basic and fundamental grouping computations. It has a position in the managed learning field and finds a lot of use in design recognition, data mining, and disruption identification. [10]. The KNN uses the Feature extraction capability to calculate distances between current data points and any new data points. [11].

#### 3.2. Logistic Regression

A true model is Logistic Regression, at its core, relies on a defined capacity to demonstrate a paired dependent variable, but it may be extended in a variety of ways. Calculated relapse is a type of relapse investigation that explores the limitations of an important model. A parallel logistic model features a dependent variable with two alternative characteristics, similar to the true/false up short catered to by a marker variable, where the two qualities are designated "0" and "1" in mathematics [12]. The calculated sigmoid capacity is frequently indicated as  $y$  [1]:

$$y = \frac{1}{1 + e^{-x}} \quad [1]$$

Where  $y$  is the yield which is reliant of information factors  $x$ .

#### 3.3. Linear Regression

Linear Regression endeavors displaying the two-factor connection by applying a linear condition on data that has been noticed. A single factor is viewed as a logical variable, and the other is viewed as a reliant variable. Equation form of Linear Regression:

$$Y = b + aX \quad [2]$$

The logical variable is  $X$ , and the dependent variable is  $Y$ . The slant of the line is  $a$ , and  $b$  is the catch [13].

### 3.4. Decision Tree

An administered characterization technique is typified by Decision Trees. The idea was inspired by the typical tree structure, which consists of a root, hubs (the points where branches split), branches, and leaves. Similarly, a Decision Tree is made up of hubs that address circles, and the fragments that connect the hubs address the branches. A Decision Tree starts at the top and works its way down, attracting people from left to right for the most part. The root hub is the point at which the tree begins. Furthermore, the information is split at the choice nodes [14].

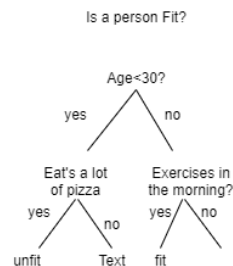


Fig: 3.1 Decision Tree

### 3.5. Gradient Boosting Classifier

Gradient boosting classifiers are a group of Machine Learning Algorithms that combine several delicate learning models to form a powerful perceptive model. When it comes to gradient boosting, decision trees

are frequently utilized. Gradient boosting has been used in a number of different disciplines. The Algorithm can look convoluted from the start, however by and large, we utilize only one predefined arrangement for order and one for relapse, which can obviously be altered dependent on your necessities. In this article, we'll center around Gradient Boosting for characterization issues. We'll begin with a gander at how the calculation functions in the background, naturally and numerically [15].

### 3.6. Light Gradient Boosting Machine

Light GBM is a quick, circulated, widely distributed, superior inclination boosting structure based on a decision tree calculation that may be used for location, characterization, and a variety of other machine learning tasks. It splits the tree leaf-wise with the best fit since it relies on decision tree calculations, unlike other boosting calculations split the tree profundity astute or level-wise rather than leaf-wise [16]. When developing in Light GBM on a similar leaf, the leaf-wise calculation can reduce more misfortune than the level-wise calculation, resulting in far greater precision than can be achieved by any of the present boosting algorithms on occasion [16].

## **CHAPTER 4**

### **Methodologies**

#### 4.1. Data Description

This investigation entails, Pima Indian Diabetes Dataset [18] was performed and a sum of 768 members is chosen matured 21 or more, all of them are women. Members of the group were

approached a survey must be completed that appeared Table 1 was created by self-arranged dependent on the requirements this could lead to diabetes.

### Pima Dataset

Parameters	Instance
Total number of participants	768
Age	21 years old or older
Gender	768
• All are female	
Pregnancies	Numeric
Glucose	In an oral glucose resistance test, plasma glucose fixing takes two hours.
Blood pressure	Blood pressure in millimeters of mercury (mm Hg)
Skin thickness	The thickness of the skin overlay over the back of the arm muscles (mm)
Insulin	Insulin levels in the blood after two hours (mu U/ml)
BMI	BMI (weight in kilograms divided by height in meters) (height in m)
Diabetespigreefunction	Function of the diabetes pedigree
Outcome	• Diabetic – 268

• Non- diabetic - 500

Table.1. Dataset describe

The dataset is a class offset type dataset with 8 credits, remembering 500 diabetes events for an aggregate of 768 gathered examples. Besides, the dataset is fragmented and some Body Mass Index (BMI) things are absent. Hazard factors in this dataset incorporate Glucose, DiabetesPedigreeFunction, BMI, Age, blood pressure, skin thickness, Insulin, and Pregnancies.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1
...	...	...	...	...	...	...	...	...	...
763	10	101	76	48	180	32.9	0.171	63	0
764	2	122	70	27	0	36.8	0.340	27	0
765	5	121	72	23	112	26.2	0.245	30	0
766	1	126	60	0	0	30.1	0.349	47	1
767	1	93	70	31	0	30.4	0.315	23	0

Fig:4.1. Dataset used in this study

## 4.2. Preprocessing of dataset

### 4.2.1. Data Analysis

For data analysis here used Exploratory Information Investigation (EDA). Exploratory information investigation (EDA) is a grounded measurable practice that gives calculated and computational apparatuses to finding examples to encourage theory improvement and refinement. These instruments and perspectives supplement the utilization of importance and speculation tests utilized in corroborative information investigation (CDA). EDA assists one with interpreting the consequences of CDA and may uncover startling or deceiving designs in the information [17].

### 4.2.2. Data Imputation

Missing data imputation is a significant undertaking in situations where it is critical to utilize all accessible data and not dispose of records with missing values. This work assesses the exhibition

of a few measurable and Machine Learning ascription techniques that were utilized to imputation repeat in patients in a broad genuine type 2 diabetes informational index.

### 4.3. Design and Implementation

For this project, Jupyter Notebook (Anaconda3) was used for implementation, while Python was used for coding. The Pima dataset was used to test machine learning techniques such as the Logistic Regression Method, Decision Tree Classifier, Linear Regression, K-Nearest Neighbors, Light Gradient Boosting Machine, and Gradient Boosting Classifier to predict diabetes. Then, as shown in fig. 4.2, compare all of these predictions from each classifier with those from other classifiers and continue the procedures to apply machine learning techniques

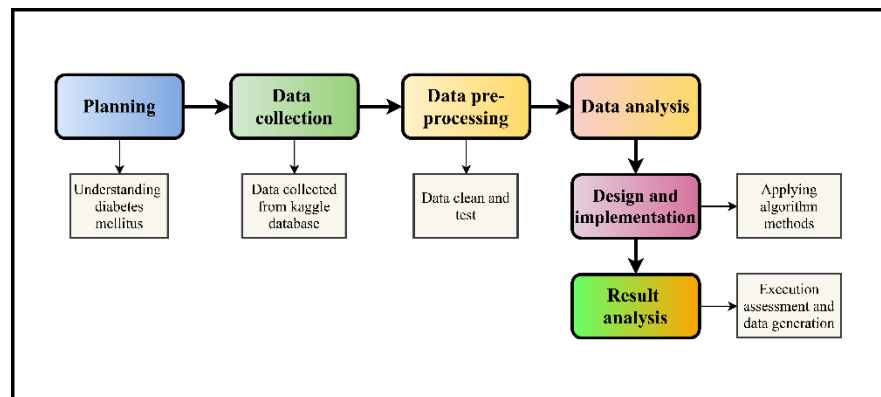


Fig:4.2. Flowchart of this Research

## CHAPTER 5

### Result & Discussion

The dataset used to predict diabetes is shown in Figure 4.1. The outcome parameter was used as a dependant variable, whereas independent variables were employed for the remaining parameter.. The diabetic line is defined by binary features. where 0 refers to non-diabetics and 1 refers to diabetics. To get started with the dataset, the entire example is divided into two halves, each with a 60:40 split for the training and test sets. To train the dataset to predict the test dataset outcomes, we used six classification techniques: Logistic Regression Method, Decision Tree



Classifier, Linear Regression, K-Nearest Neighbors, Light Gradient Boosting Machine, and Gradient Boosting Classifier, as indicated in table 5.1.

Pima Dataset	Logistic Regression			Decision Tree Classifier			K-Nearest Neighbors			Light Gradient Boosting Machine			Gradient Boosting Classifier		
	0	1		0	1		0	1		0	1		0	1	
	0	267	33	0	233	67	0	249	51	0	251	49	0	257	43
	1	67	94	1	63	98	1	73	88	1	69	92	1	68	93

Table:2. Different ways hold a confusion matrix

The next metric is as follows condition 3-6 may to figure out from the obtained disarray grids. TN, FP, FN, and TP were the results of these matrices (TP). In the dataset, the TN is greater than the TP. In this manner, every one of the techniques are giving acceptable outcomes. To track down the specific exactness of each technique the accompanying measures have been determined by given formula [ [18], [19], [20]]:

$$\text{Accuracy Rate} = \frac{TP+TN}{TP+TN+FN+FP} \quad [3]$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad [4]$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad [5]$$

$$\text{F-Measure} = \frac{2 * (\text{Precision} * \text{Sensitivity})}{\text{Precision} + \text{sensitivity}} \quad [6]$$

Every one of the model exhibitions are estimated against Accuracy, Recall, precision, and F-measure which is appeared in Table 5.2. But only the Linear Regression model gave accuracy.

	Logistic Regression	Decision Tree	Linear Regression	K-Nearest Neighbors	Light Gradient Boosting Machine	Gradient Boosting Classifier
Accuracy	0.78	0.72	0.32	0.73	0.74	0.76
Recall	0.58	0.61		0.55	0.57	0.58
Precision	0.74	0.59		0.63	0.65	0.68
F-measure	0.65	0.60		0.59	0.61	0.63

Table.3. Qualities for various measures for various classification techniques

Table 3 also shows that the Logistic Regression classifier has the highest accuracy (78%), precision, and F-measure of all the methods, indicating that it is the best for the Pima dataset.

## CHAPTER 6

### Conclusion

#### 5.1 Findings

One of the world's most pressing medical issues is recognizing the dangers of diabetes in it is still in its early phases. The goal of this study is to develop a system that can assess the risk of diabetes type 2. Six classifiers based on machine learnings algorithms have been put to the test. in this article, and different factual measures were used to compare and contrast their outcomes. Tests were run using a dataset compiled from online and disconnected questionnaires that included eight diabetes-related questions. The exactness of the Logistic Regression of Pima dataset is 78 percent, which is the highest among the others. Every one of the six machine learning algorithms used produced excellent results for various measures such as accuracy, recall, and so on.

## 5.2 Recommendation for Future works

In order to continuously prepare and streamline our suggested model, it is critical to obtain the medical clinic's real and most recent patients' information for future work. The dataset's size should be sufficient for further preparation. In order to investigate DM, some high-level computations and models need be used. Reviewing estimating fundamentals is a better way to go. This will aid in slowing the progression of diabetes and, as a result, lowering the risk of developing DM.

## References

- [1] Neha Prerna Tigga , Prediction of Type 2 Diabetes using Machine Learning, Mesra, Ranchi, India : Procedia Computer Science, 2020.
- [2] M. S. V. U. R. B. A. J. S. R. J. P. P. Y. C. S. D. V. K. Datta, Prevalence of diabetes and prediabetes, urban and rural India: Indian Council of Medical Research–INdia DIABetes, 2011.
- [3] American Diabetes Association, Gestational Diabetes Mellitus, Alexandria: Diabetes Care, suppl. American Diabetes Associaton, 2004.
- [4] WS Weintraub, SR Daniels, LE Burke, BA Franklin , "Value of primordial and primary prevention for cardiovascular disease: a policy statement," *Journal of the American Heart Association*, 2020.

- [5] JL Murphy, EA Girot, "The importance of nutrition, diet and lifestyle advice for cancer survivors—the role of nursing staff and," *Journal of Clinical Nursing*, 2013.
- [6] Wu, H., Yang S., Huang, Z., He, J., Wang, X., Type 2 diabetes mellitus prediction model based on data mining, *Informatics in Medicine Unlocked*, 2018.
- [7] Swapna, G., Vinayakumar R., Soman K. P. , "Diabetes detection using deep learning algorithms," *ICT Express* , 2018.
- [8] D. S. D. S. Sisodia, "Prediction of diabetes using classification algorithms," *Procedia computer science*, 2018.
- [9] N. M. R. Nai-arun, "Comparison of classifiers for the risk of diabetes prediction," *Procedia Computer Science*, 2015.
- [10] Oliver Kramer, "K-Nearest Neighbors," *Springer*, 2013.
- [11] Leif E. Peterson , "K-nearest neighbor," *scholarpedia*, 2009.
- [12] L. G. Grimm & P. R. Yarnold , "Logistic regression," *American Psychological Association*, 1995.
- [13] Odd O. Aalen , "A linear regression model for the analysis of life times," *Statistics in medicine*, 1989.
- [14] Z. a. Y. L. Yan-yan SONG1, "Decision tree methods: applications for classification and prediction," *Shanghai archives of psychiatry*, 2015.
- [15] Alexey Natekin1\* and Alois Knoll, "Gradient boosting machines, a tutorial," *Frontiers in neurobotics* , 2013.
- [16] X. M. L. W. F. Z. X. Y. W. Z. -. A. W. J Fan, "Light Gradient Boosting Machine: An efficient soft computing model for estimating daily reference evapotranspiration," *Elsevier*, 2019.
- [17] Behrens, J. T. , "Principles and procedures of exploratory data analysis," *Psychological Methods*, 1997.
- [18] A. B. R. K. H. K. Y. A. K. O. P. P. J. H. H. U. O. Käräjämäki, "Non alcoholic fatty liver disease with and without metabolic syndrome: different long-term outcomes.," *Metabolism*, 2017.
- [19] M. G. S. H. M. S. K. S. M. V. A. G. Y. C. M. P. T. D. M. Gurka, "Independent associations between a metabolic syndrome severity score and future diabetes by sex and race:," *Diabetologia*, 2017.
- [20] M. K. J. S. A. K. T. P. P. L. A. J. C. F. S. M. K. L. B. M. Laakso, "The Metabolic Syndrome in Men study: a resource for studies of metabolic and cardiovascular diseases," *Journal of lipid research*, 2017.

