ISSN NO: 2770-4491
Date of Publication:04-08-2023

Developing a ML-Based Model for Detecting Diabetes

Ban Hamid Ali,
Faculty of Arts and Science, Computer Science Department, American University of Culture and Education, Beirut, Lebanon
ban_1985@yahoo.com, bha013@auceonline.com
Dr. Mahmoud koabaz,
Faculty of Arts and Science, Computer Science Department, Assistant Professor, American University of Culture and Education, Beirut, Lebanon
mahmoudkoabaz@auce.edu.lb, m.koabaz@gmail.com

Abstract: As technology progresses, data grows exponentially and is produced quickly from a wide variety of sources. Data storage, analysis, and interpretation become challenging due to data's variety and complexity. Despite the fact that we can now store such enormous amounts of data, much progress has to be made in terms of analyzing and making sense of that data. Better clinical routes to detect and predict diabetes at early stages are essential for reducing complications and delaying the onset of diabetes.
The worldwide rise in the prevalence of diabetes is one of the most concerning trends in modern medicine. A high blood glucose level, caused by insufficient insulin synthesis or improper insulin response by body cells, characterizes the metabolic illnesses known together as diabetes mellitus. Organs as diverse as the brain, nervous system, heart, kidney, eyes, skin, and limbs may all suffer harm from diabetes mellitus.
In this study, we used machine learning to differentiate between type 1 and type 2 diabetes trends

Keyword: analysis, techniques, classification, machine learning

In this study, we used machine learning to differentiate between type 1 and type 2 diabetes trends. Our initial findings were less than ideal, but by using a data rebalancing strategy, we were able to get a high accuracy of the model (about 0.85, with 0.87 as Precision and Recall Value for Type1 and 0.83 for Type2). This is a novel paradigm that when compared to other efforts, turns out to be more effective and reliable when applied to Type1 and Type2 diabetes.

Introduction

Hence the medical domain researchers depending on IT solutions to eliminate such errors to possible extend and constantly trying to get benefits in the specific area like Diabetes Retinopathy which lies within the critical application area. Predominantly the computer-vision-oriented image analyzers are used to give physicians a quicker, accurate, and target-aligned second opinion on investigations made initially
The ophthalmologists struggle usually with noisy retinal images for deciding the presence of abnormalities which are highly influenced by hypertension, cardio vascular disorder, Glaucoma, and primarily by age. So far, the efforts in this direction yield early detection of accuracy relatively poor even less than or around 60%. The most frequent defect in the eyes is due to diabetic retinopathy and this defect occurs in the absence of attention for many years [1]
Studies have shown that the selective application of routines for exploring improvements of diagnostic accuracy by fixing a carefully designed framework. In this digital era, data is center of the knowledge, economy, and technology. Data has established an important emergent scientific paradigm driving research evolution in such disciplines as statistics, computing science, social science, and intelligence science [3].
the characteristics of data like volume, velocity, veracity, variability, validity, etc. require advanced computational intelligence to analyze the data [4].

Although there are some traditional/conventional techniques which are being used for the analysis of data. But they fail to deliver owing to the certain issues regarding data like high velocity, high volume, high variety,

ISSN NO: 2770-4491
Date of Publication:04-08-2023

etc. The data from varied sources goes beyond the analytical capabilities of humans and traditional database management systems to process this huge volume of data.

Conventional/traditional methods are not able to handle such voluminous data in order extract the information from it for decision making. Conventional methods require some advanced analytical tools and techniques to handle data so that this raw data can be converted into information and then knowledge to provide solutions to real-life problems [7].

Data analytics also makes use of statistical analysis techniques. Analysis of company data using statistical methods in order to provide useful insights and enhance strategic decision making. Information or knowledge may be extracted from raw data by statistical methods, but this requires human analysis and interpretation

The capacity of a human specialist in a certain subject to do conventional or statistical analysis and generate actionable conclusions from the data is crucial to the success of such an analysis. Such approaches, however, are impractical because they are excessively time-consuming, expensive, and reliant on the expert's knowledge, insight, and analysis [8].

Literature Review

Introduction

In recent years, data mining for predicting the onset of diabetes complications has become more popular. Many different toolkits and algorithms have been investigated and developed by researchers. Using these frameworks and techniques, the vast potential of this area of study is made clear. In this chapter, we take a look at a few key studies that have had, or will have, major effects on diabetes detection and diagnosis.

Data Analytics

Data analytics is the process of analyzing data by using distributed, highly scalable systems and frameworks that can process massive amounts of data from many sources [7]. Data management is the process of overseeing all aspects of the data-gathering, processing, storing, analyzing, and reporting phases of the data lifecycle to ensure useful information is produced and available for use in making decisions. [8].

Data analytics may help businesses optimize their procedures by learning more about their customers' preferences and habits. support systems in the organizations. Different techniques with different goals are being explored in data analytics based upon the outcome of process, but these techniques are being summarized into four types as discussed below [11].

1- Descriptive: Descriptive analytics basically answers the questions. It manipulates a huge amount of data by using data aggregation and data mining techniques in order to provide insights into the past.

2. Diagnostic: It is used to determine why something happened in the past. It uses a combination of techniques like drill-down, data discovery, data mining, and correlations.

3- Predictive: It uses the advantage of both descriptive and diagnostic analytics to predict what might happen in the future. It is based on statistical modeling along with machine learning and deep learning to predict future trends, which makes it a valuable tool for forecasting

4- Prescriptive: The purpose of prescriptive analytics is to use optimization and simulation algorithms for the possible outcome so that what action to take in order eliminates future problems or take full advantage of promising trends

ISSN NO: 2770-4491
Date of Publication:04-08-2023

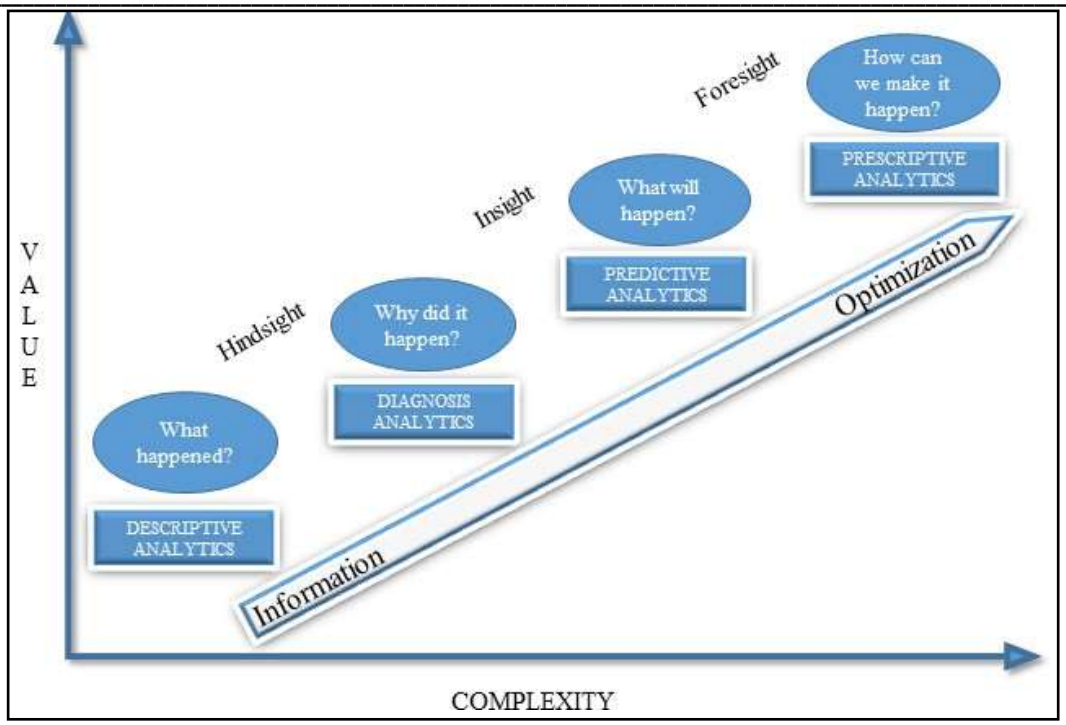


Figure. 2.1: Data Analytics at Different Phases

Data Mining, Machine Learning, and Deep Learning are three promising new areas of research in healthcare analytics. Healthcare data comes from many different places and in many different formats. Integrating and bringing health data into a common platform for further analysis requires the use of sophisticated tools and processes in order to create meaningful information and, ultimately, knowledge.
Predicting clinical paths, improving novel medical treatments, and extracting hidden information from the massive number of patient data and records are just a few examples of the many ways in which machine learning methods contribute to health-related issues [10]. There are just a few tangible advantages of using ML in healthcare, and they are as follows:

- Diabetes, cardiovascular disease, malaria diagnosis, breast cancer, kidney cancer, Alzheimer's disease, Parkinson's disease, and many more illnesses may be predicted with the use of ML.
- In order to aid in the detection of breast cancer, Google has recently created an algorithm.
- According to research conducted at Stanford University, a deep learning system may detect skin cancer. [23].
- By using optimization strategies, the Quotient healthcare application helps bring down the price of EMR. [24].
- CIOX Health investigates machine learning strategies for enhancing the healthcare system via the sharing of health data among professionals in the field.
- Path AI's employed machine learning to help pathologists to identify new treatments and accurate diagnosis.
- Artificial intelligence-based Berg's Massachusetts-based Interrogative Biology platform uses machine learning for disease mapping and medication development in cancer, neuroscience, and other uncommon disorders.
- The MD Insider Platform explores machine learning techniques to match patients with healthcare providers in order to provide better quality of service in healthcare.
- Beta Bionics is a smart wearable device focused on the "bionic" pancreas system by calling 'iLet' which manages the blood sugar levels in diabetic patients around the clock.

ISSN NO: 2770-4491
Date of Publication:04-08-2023

Machine learning and deep learning paradigms in healthcare are being widely used and are helping patients and healthcare providers in different ways. The performance evaluation of machine learning based automatic systems for prediction, detection, prognosis, and diagnosis have proved to be realistic and nearly equivalent to that of healthcare providers [26]. ML and DL tools and techniques whenever implemented in healthcare industry have yielded better results [27]. E.g., in radiology, ML and DL techniques detect and predict complex patterns automatically, and helps radiologists to make intelligent decisions reviewing images such as CT, PET images, MRI, and radiology reports [28].

Diabetes Mellitus Disease

The rising incidence of diabetes throughout the globe is a huge public health problem. "Diabetes mellitus is the deadliest and most common of a group of metabolic disorders in which insulin production is inadequate or insulin is not used efficiently by the body." [21]. High blood glucose levels, either by insufficient insulin synthesis or improper insulin response by body cells, characterise the metabolic illnesses known together as diabetes mellitus [22].

Different forms of diabetes include [31].

1. TYPE 1: Also known as juvenile diabetes or insulin-dependent diabetes and is caused due to autoimmune dysfunction. The disease can develop at any stage but most frequently occurs in children and adolescents.
2. TYPE 2: It is also called as non-insulin-dependent diabetes. It is the most common type of diabetes and is characterized by inadequate production of insulin in the body. It affects all the age groups and patients often exhibit obesity, overweight, urination, etc. which correlates with the presence of insulin resistance.
3. TYPE 3: Also known as gestational diabetes in which hyperglycemia condition occurs in women during pregnancy. This type of diabetes increases the likelihood of type-2 diabetes in the mother and fetus.
4. TYPE 4: Pre-diabetes is characterized by elevated blood sugar levels and increased insulin production and may result from hereditary problems, pharmacological side effects, or an increase in other hormone levels. [33].

Diet and Physical Exercise Plans for T2DM

Recent experimental research in the medical sciences has shown the importance of nutrition and exercise in preventing, delaying, and treating diabetes, especially Type-2 Diabetes Mellitus. Various chronic illnesses have been linked to dietary and activity patterns, according to the existing research.

It means that precautions and healthy lifestyle are far better than the use of medication (insulin and drugs) to control this disease. The various factors that affect diabetes is shown in Figure 2.2.

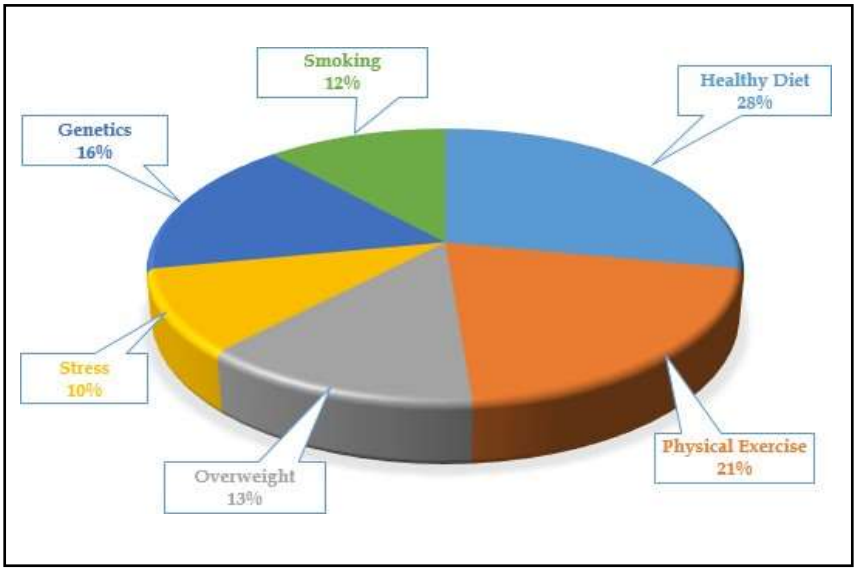


Figure 2.2: Contributing Factors towards T2DM

Analysis of Preprocessing Methods for Diabetes Mellitus

In order to offer an efficient pre-processing strategy, Asgarnezhad et al. [11] combined attribute subset selection methods with missing value replacement. The most widely utilised database for studying diabetes mellitus is put to good use here. Experimental findings and diabetes mellitus prediction demonstrate that the suggested method may improve the performance of the used classifier, outperforming standard methods in terms of precision and accuracy.

Wosiak and Karbowiak [19] Developed a pre-processing compensation approach to correctly classify medical datasets with significant imbalances. Classification with a certain preprocessing strategy may improve results for particular datasets. In datasets with a lot of characteristics or a complicated distribution, the number of accurate predictions will drop no matter what classification algorithm or resampling methodology is used.

Albayrak et al., [14] used a missing data imputation strategy based on Maximum Likelihood Estimation (MLE) and clustering algorithms. Missing data patterns such as Missing Not at Random (MNAR), Missing at Random (MAR), and Missing Completely at Random (MCAR) are accommodated by creating a new dataset in the first phase.

Almuhaideb and Menai [18] utilized pre-processing strategy for selecting an optimal feature combination, which has a major effect on data categorization capability. Pre-processing activities such as dealing with missing data, selecting subsets of characteristics, and discretizing numeric attributes are taken into account. In this example, we use an Ant Colony Optimization Algorithm to the problem of categorization. Experimenting with 25 real-world medical datasets yields a relative improvement in predicting accuracy of over 60%.

Analysis of Feature Selection Techniques

Sneha and Gangil et al, developed a prediction method using Machine Learning, and identified the best classifier for getting results that are very near to clinical outcomes. This suggested study uses Predictive analysis to zero down on attributes selection, a technique employed in the early identification of Diabetes Mellitus. Experiments on data from people with diabetes demonstrate that both the Decision Tree algorithm and Random Forest approaches provide findings with a specificity of roughly 98.20 percent.

Analysis of Classification Techniques

Kumar Dewangan and Agrawal et al, used two approaches to create a model for an ensemble. Multilayer Perceptron and Bayesian classification are the two techniques used. Diagnosing diabetes mellitus may be evaluated in terms of its specificity, sensitivity, and accuracy. This model uses six characteristics to achieve an accuracy of around 81.89 percent, with a sensitivity of about 64 percent and a specificity of about 90 percent.

Summary

This chapter describes a review of the on-going research and also focuses on recent developments in Machine Learning which have made significant impacts on the detection and diagnosis of diabetes. This review gives an elaborate examination of the procedures, methodologies, qualities, and undersized comings.

The Methodlogy

Introduction

In the past decade, introduction to machine learning has transformed several industries, including healthcare, costumer services, social media, transportation, education, e-commerce, manufacturing, etc. Machine learning paradigms play an important role in designing and developing realistic frameworks for healthcare systems. A trend in healthcare analytics is taking advantage of AI-based techniques by using intelligent software and hardware in order to find out meaningful information from large and complex data repositories.

Predictive modelling, in which these methods have been used, allows for the creation of one or more models with foresight capabilities and trends. Data about one's way of life may also be fed into a system trained using machine learning and deep learning methods. For instance, these days' people may snap images of their food and have it identified by a prediction model like a deep neural network [20]; this is very useful in the medical field

Exploratory Data Analysis

Exploratory Data Analysis (EDA) is considered as a major step in machine learning process and are being used to improve the quality assessment of data [24]. The real-world data can be typically unstructured, inconsistent, and generally being generated from different sources that needs to be preprocessed to derive

ISSN NO: 2770-4491
Date of Publication:04-08-2023

insights before it can be used to solve a real-world problem. This is where EDA techniques came to rescue, it helps to integrate, clean, transform, encode, format, and organize the raw data, thereby making it ready-to-go for building machine learning based frameworks [25].

EDA is not a single-step process; it involves different tasks as shown in Figure 3.1. The methods of Exploratory Data Analysis are data cleaning, data scaling, and data transformation.

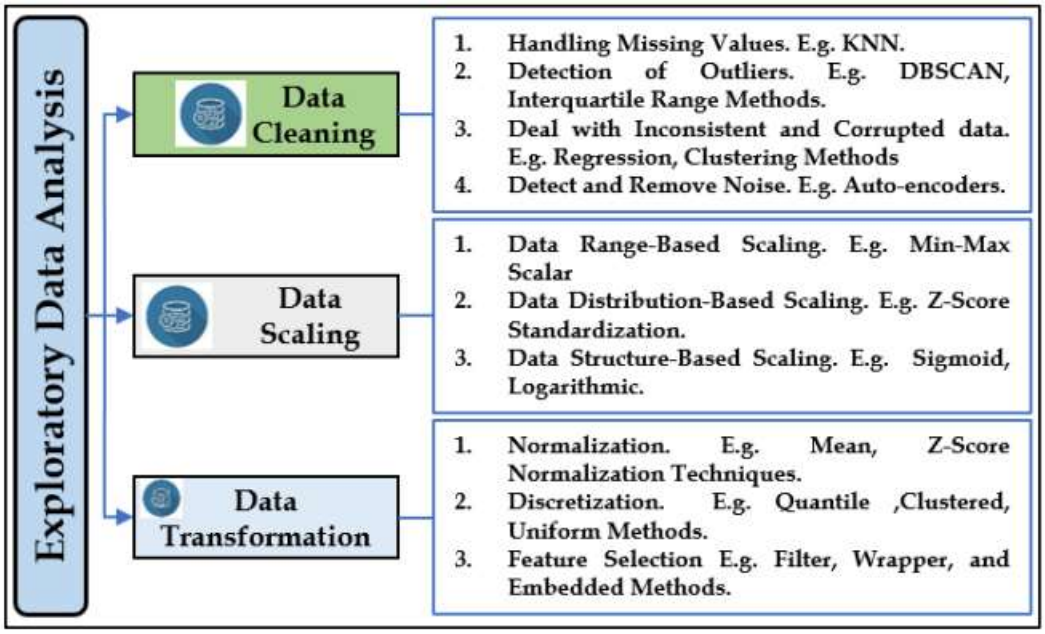


Figure 3.1: Exploratory Data Analysis

The EDA methods used to analyze dataset and summarize their main characteristics even visualization methods are discussed in detail as**:**

Data Cleaning: Handling of missing values, detection of outliers, and dealing with inconsistent, corrupted, and noisy data can be performed using some statistical tools of ML libraries.

Data scaling: It can be employed for data distribution and scaling purposes to draw the proper format before developing the ML/EL models.

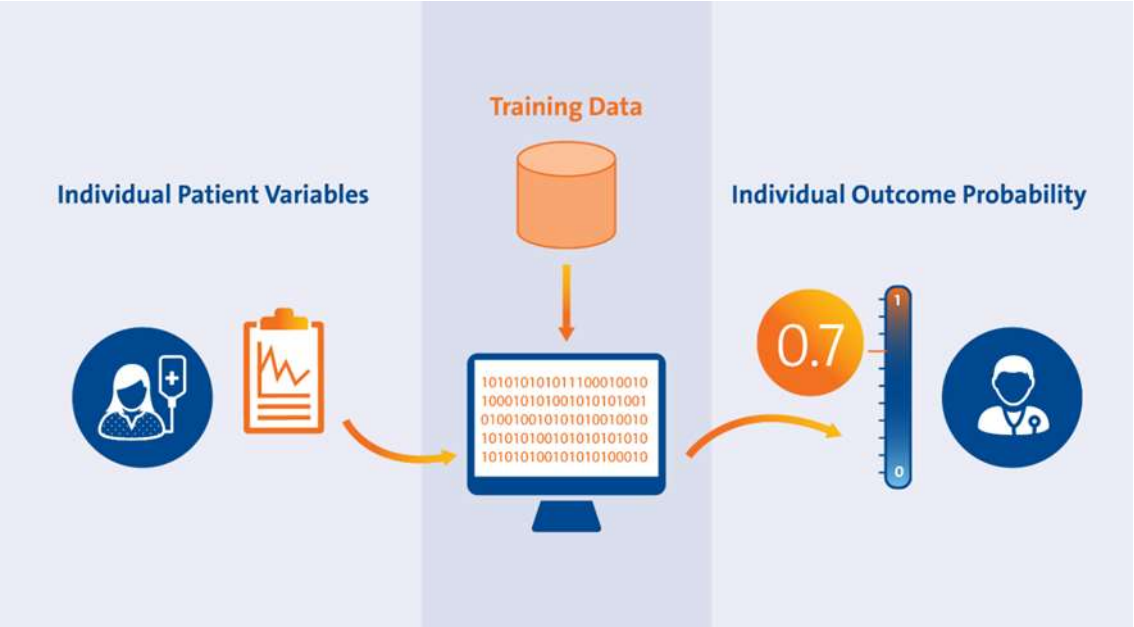
Data Transformation: In transformation method, the data is normalized and discretized based on data distribution of samples in dataset. Also, contribution of features/parameters towards the disease are measured to enhance model accuracy.

THE IMPLEMNTATION

Introduction

Machine learning prediction it is the output of algorithms trained on historical data. After training, the algorithm generates potential values for unknown variables in each reading or record of new examples that have not been trained by the model.

ISSN NO: 2770-4491
Date of Publication:04-08-2023



Prediction and classification are two sides of the same coin, in which the model in classification tries to predict the correct label of the input samples after training the model on the entire training sample and then it is evaluated later on the test samples to test the accuracy of the model and discover its ability to correctly classify. So classification includes prediction within it.

The Dataset

When searching on the web for data sets on diabetes, we quickly find that famous data set, Pima, which classifies the patient as a diabetic or a healthy person, but does not contain readings about classifying the disease as whether it types one or type two.

And when we searched deeply, we found only a small data set that meets the purpose, which is Data Shanghai, which contains only dozens of readings about the first and second types of diabetes.

Data Description

The data consists of 125 lines, 109 of which belong to the second pattern, and only 16 belong to the first pattern.

It also contains the following 26 columns:

1 - Gender (Female=1, Male=2)
2 - Age (years)
3 - Height (m)
4 - Weight (kg)
5 - BMI (kg/m2)
6 - Smoking History (pack year)
7 - Alcohol Drinking History (drinker/non-drinker)
8 -Type of Diabetes (Type1 or Type2)
9 - Duration of diabetes (years)
10 - Fasting Plasma Glucose (mg/dl)
11 - 2-hour Postprandial Plasma Glucose (mg/dl)
12 - Fasting C-peptide (nmol/L)
13 - 2-hour Postprandial C-peptide (nmol/L)
14 - Fasting Insulin (pmol/L)
15 - 2-hour Postprandial insulin (pmol/L)
16 - HbA1c (mmol/mol)
17 - Glycated Albumin (%)
18 - Total Cholesterol (mmol/L)

ISSN NO: 2770-4491
Date of Publication:04-08-2023

- Triglyceride (mmol/L)
- High-Density Lipoprotein Cholesterol (mmol/L)
- Low-Density Lipoprotein Cholesterol (mmol/L)
- Creatinine (umol/L)
- Estimated Glomerular Filtration Rate (ml/min/1.73m2)
- Uric Acid (mmol/L)
- Blood Urea Nitrogen (mmol/L)
- Hypoglycemia (yes/no)

Importing Dataset

The available data files are in the form of Excel files, so they will be combined with each other and then called in Jupiter Notebook.

```
import pandas as pd
from pandas import ExcelWriter
from pandas import ExcelFile

df=pd.read_excel("Diabetes Type1 Type2.xlsx")

df
```

	Gender (Female=1, Male=2)	Age (years)	Height (m)	Weight (kg)	BMI (kg/m2)	Smoking History (pack year)	Alcohol Drinking History (drinker/non-drinker)	Type of Diabetes	Duration of diabetes (years)	Fasting Plasma Glucose (mg/dl)	...	Glycated Albumin (%)	Total Cholesterol (mmol/L)	Triglyceride (mmol/L)	High-Density Lipoprotein Cholesterol (mmol/L)
0	2	57	1.69	67.4	23.600000	0.0	non-drinker	T2DM	25.000000	82.44	...	14.9	4.07	0.61	2.28
1	1	69	1.45	55.8	26.540000	0.0	non-drinker	T2DM	14.000000	137.7	...	15.3	4.36	0.89	1.36
2	1	69	1.45	53.2	25.300000	0.0	non-drinker	T2DM	14.000000	138.42	...	17	4.28	0.93	1.13
3	2	57	1.70	70.0	24.221453	30.0	drinker	T2DM	10.000000	158.4	...	29.8	5.31	1.32	1.33
4	1	58	1.64	62.5	23.237656	0.0	non-drinker	T2DM	22.000000	194.4	...	19.5	4.73	1.73	0.91

Type of Diabetes	Duration of diabetes (years)	Fasting Plasma Glucose (mg/dl)	...	Glycated Albumin (%)	Total Cholesterol (mmol/L)	Triglyceride (mmol/L)	High-Density Lipoprotein Cholesterol (mmol/L)	Low-Density Lipoprotein Cholesterol (mmol/L)	Creatinine (umol/L)	Estimated Glomerular Filtration Rate (ml/min/1.73m2)	Uric Acid (mmol/L)	Blood Urea Nitrogen (mmol/L)	Hypoglycemia (yes/no)
T2DM	25.000000	82.44	...	14.9	4.07	0.61	2.28	1.86	105	68	245	4.9	yes
T2DM	14.000000	137.7	...	15.3	4.36	0.89	1.36	2.89	41	101	337	5.6	yes
T2DM	14.000000	138.42	...	17	4.28	0.93	1.13	2.94	39	103	262	3.7	no
T2DM	10.000000	158.4	...	29.8	5.31	1.32	1.33	3.68	71.5	105	475.6	4.74	no
T2DM	22.000000	194.4	...	19.5	4.73	1.73	0.91	2.95	36.9	166	490.05	5.54	no

ISSN NO: 2770-4491
Date of Publication:04-08-2023

```
# Importing Libraries

import warnings
warnings.filterwarnings('ignore')
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import LabelEncoder, OneHotEncoder, StandardScaler
from sklearn.model_selection import train_test_split
from sklearn.model_selection import GridSearchCV
from sklearn.linear_model import LogisticRegression
from xgboost import XGBClassifier
from sklearn.metrics import mean_absolute_error, accuracy_score, classificati
```

Conclusion

In this thesis, we were able to classify the types of diabetes, into a Type1 and a Type2, using machine learning algorithms, after selecting the relevant data set and performing some necessary processing, where several machine learning models were applied, at first we got good results, but after Using the data rebalancing technique we got amazing results

References

[1] “International federation of diabetes,” https://www.idf.org, accessed: 201901-12.
[2] T. Y. Wong, J. Sun, R. Kawasaki, P. Ruamviboonsuk, N. Gupta, V. C. Lansingh, M. Maia, W. Mathenge, S. Moreker, M. M. Muqit et al., “Guidelines on diabetic eye care: The international council of ophthalmology recommendations for screening, follow-up, referral, and treatment based on resource settings,” Ophthalmology.
[4] S. Lin, P. Ramulu, E. L. Lamoureux, and C. Sabanayagam, “Addressing risk factors, screening, and preventative treatment for diabetic retinopathy in developing countries: a review,” Clinical & experimental ophthalmology, vol. 44, no. 4, pp. 300–320, 2016.
[7] M. W. Stewart, “Treatment of diabetic retinopathy: recent advances and unresolved challenges,” World journal of diabetes, vol. 7, no. 16, p. 333, 2016.
[8] T. Das, R. n, K. Ramasamy, and P. K. Rani, “Telemedicine in diabetic retinopathy: current status and future directions,” Middle East African journal of ophthalmology, vol. 22, no. 2, p. 174, 2015.
[11] M. Abra`moff, J. Reinhardt, S. Russell, J. Folk, V. Mahajan, M. Niemeijer, and G. Quellec, “Automated early detection of diabetic retinopathy,” Ophthalmology, vol. 117, no. 6, pp. 1147–1154, 2010.
[14] D. DeBuc, “A review of algorithms for segmentation of retinal image data using optical coherence tomography,” Image Segmentation,”(InTech, 2011), 2011.
[19] “Diabetic retinopathy,” https://www.aoa.org, 2019, accessed: 2019-01-22.
[21] W. Bu, X. Wu, X. Chen, B. Dai, and Y. Zheng, “Hierarchical detection of hard exudates in color retinal images,” Journal of Software, vol. 8, no. 11, pp. 2723–2732, 2013.
[22] K. M. Adal, P. G. Van Etten, J. P. Martinez, K. W. Rouwen, K. A. Vermeer, and L. J. van Vliet, “An automated system for the detection and classification of retinal changes due to red lesions in longitudinal fundus images,” IEEE Transactions on Biomedical Engineering, vol. 65, no. 6, pp. 1382–1390, 2018.
[24] A. Sopharak, B. Uyyanonvara, and S. Barman, “Automated microaneurysm detection algorithms applied to diabetic retinopathy retinal images,” Maejo International Journal of Science and Technology, vol. 7, no. 2, p. 294, 2013.
[25] K. Ram, G. Joshi, and J. Sivaswamy, “A successive clutter-rejection-based approach for early detection of diabetic retinopathy,” Biomedical Engineering, IEEE Transactions on, vol. 58, no. 3, pp. 664–673, 2011.
[26] R. K. Pappuru, L. Ribeiro, C. Lobo, D. Alves, and J. Cunha-Vaz, “Microaneurysm turnover is a predictor of diabetic retinopathy progression,” British Journal of Ophthalmology, vol. 103, no. 2, pp. 222–226, 2019.

ISSN NO: 2770-4491
Date of Publication:04-08-2023

[28] M. Garc´ıa, C. S´anchez, M. Lo´pez, D. Aba´solo, and R. Hornero, "Neural network based detection of hard exudates in retinal images," COMPUT METH PROG BIO, vol. 93, no. 1, pp. 9–19, 2009.

[31] G. Quellec, M. Lamard, B. Cochener, C. Roux, G. Cazuguel, E. Decenciere, B. Lay, and P. Massin, "A general framework for detecting diabetic retinopathy lesions in eye fundus images," in Computer-Based Medical Systems (CBMS), 2012 25th International Symposium on. IEEE, 2012, pp. 1–6.

[32] E. Decenci`ere, G. Cazuguel, X. Zhang, G. Thibault, J.-C. Klein, F. Meyer, B. Marcotegui, G. Quellec, M. Lamard, R. Danno et al., "Teleophta: Machine learning and image processing methods for teleophthalmology," Irbm, vol. 34, no. 2, pp. 196–203, 2013

[33] A. Hoover, "Stare database," Available: Available: http://www. ces. clemson. edu/˜ ahoover/stare, 1975