The Prediction of Diabetes

Alessandro Massaro
LUM University Giuseppe Degennaro; Lum Enterprise s.r.l. https://orcid.org/0000-0003-1744-783X
Nicola Magaletti
LUM Enterprise s.r.l.
Gabriele Cosoli
LUM Enterprise s.r.l.
Vito Giardinelli
LUM Enterprise s.r.l.
Angelo Leogrande (leogrande.cultore@lum.it)
LUM University Giuseppe Degennaro; LUM Enterprise s.r.l. https://orcid.org/0000-0003-1381-4006

Research Article

Keywords: Machine Learning, Clusterization, Elbow Method, Prediction, Correlation Matrix, Principal Component Analysis, Binary and non-Binary regression models

Posted Date: June 13th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1753046/v1

Alessandro Massaro[1], Nicola Magaletti[2], Gabriele Cosoli[3], Vito O.M. Giardinelli[4], Angelo Leogrande[5]

The Prediction of Diabetes

Abstract

The following article presents an analysis of the determinants of diabetes using a dataset containing the surveys of 2000 patients from the Frankfurt Hospital in Germany. The data were analyzed using the following models, namely: Tobit, Probit, Logit, Multinomial Logit, OLS, WLS with heteroskedasticity. The results show that the presence of diabetes is positively associated with "*Pregnancies*", "*Glucose*", "*BMI*", "*Diabetes Pedigree Function*", "*Age*" and negatively associated with "*Blood Pressure*". A cluster analysis is realized using the fuzzy c-Means algorithm optimized with the Elbow method and three clusters were found. Finally a confrontation among eight different machine learning algorithms is realized to select the best performing algorithm to predict the probability of patients to develop diabetes.

Keywords: Machine Learning, Clusterization, Elbow Method, Prediction, Correlation Matrix, Principal Component Analysis, Binary and non-Binary regression models.

1. Introduction-Research Question

Diabetes is one the most relevant contemporary disease. 10% of the US population has diabetes. 84 million Americans are prediabetes and are 70% likely to develop type 2 diabetes in the absence of health interventions [1]. Furthermore, diabetes is one of the most costly disease for national health systems. The combination of the two factors i.e. the large percentage of world population that is affected by diabetes, and the high costs of diabetes for national health systems create the urgency for a political, institutional response that is either oriented to promote individual and social health either able to reduce the cost of diabetes treatments for patients. The change of individual behaviours, the understanding of the co-cause of diabetes, and even the research and development in the pharmacological sectors are able to promote a deeper understanding of diabetes and the methodology to promote a better health at individual and social level and a reduction in public costs for the care of diabetes patients. In our analysis we have used a public databased that consider a set of 2000 patients from a Hospital in Frankfurt that is on Kaggle. Our contribution try to investigate the presence of co-cause of diabetes trying to understand the presence of significant association between diabetes on ne side and other individual health conditions on the other side. Finally, based on the proposed variable we have estimated the probability that the n-patient could have or not diabetes based on the analyzed datasets. Our research question is based on the necessity to investigate solid relationship between

[1] Professor at Lum University Giuseppe Degennaro, and Chief Research Officer-CRO at Lum Enterprise s.r.l. Email: massaro@lum.it. Strada Statale 100 km 18, 70010 Casamassima BA, Puglia, Italy, European Union.
[2] Chief Operation Officer-COO and Senior Researcher at Lum Enterprise s.r.l. Email: magaletti@lumenterprise.it. Strada Statale 100 km 18, 70010 Casamassima BA, Puglia, Italy, European Union
[3] Senior IT Specialist and Solutions Architects and Researcher at Lum Enterprise s.r.l. Email: cosoli@lumenterprise.it. Strada Statale 100 km 18, 70010 Casamassima BA, Puglia, Italy, European Union.
[4] Business Developer and Researcher at Lum Enterprise s.r.l. Email: Giardinelli@lumenterprise.it Strada Statale 100 km 18, 70010 Casamassima BA, Puglia, Italy, European Union.
[5]Assistant Professor at Lum University Giuseppe Degennaro and Researcher at Lum Enterprise s.r.l. Email: leogrande.cultore@lum.it. Strada Statale 100 km 18, 70010 Casamassima BA, Puglia, Italy, European Union

diabetes as investigated variables and explicable variables in order to predict the probability for patients to develop diabetes based on the analyzed data.
The article continues as follows: the second paragraph contains the methodology, the third paragraphs presents the a non-critical but just informative literature review to introduce the topic, the fourth paragraph contains the data description, correlation matrix and principal component analysis, the fifth paragraph analyses the results of the regression analysis, the sixth paragraph shows the results of a clusterization with fuzzy c-Means optimized with the Elbow Method, the seventh paragraph presents the results of the confrontation among different machine learning algorithm oriented to the prediction of the probability to develop diabetes, the eight paragraphs concludes.

2. Methodological section

The article present a set of metric analysis that have the goal to set the best set of variables to generate a model oriented to predict the probability of a patient to develop diabetes. In this sense the methodology presents can be divided in two different phases:

- First phase: in which a set of complex metric analysis is realized to find the best variable for the prediction. Specifically we use a set of complex metric analysis to verify the presence of significant relationships among the variables of the datasets. The main goal of this phase is to verify the presence of variables that are associated to diabetes in the sense of correlation, with PCA and as a consequence of the regression analysis. Once the variables are found, and after checking their consistence and metric appropriateness, we build a model that in the third phase has been used to the prediction of the probability of patients to develop diabetes.
- Second phase: in this phase a cluster analysis has been realized with the fuzzy c-Means algorithm to verify the presence of relationships among each cluster. The main goal of this phase is twofold: firstly the dataset is divided in a series of clusters and secondly after having computed the percentage of diabetes patient in respect to the total number of patient, we try to verify the presence of relationships among the variables in each cluster.
- Third phase: is based on a comparison among different machine learning algorithms to predict the probability that a patient has to develop diabetes. This analysis is based on the valuation of the predictive performance of machine learning algorithms based on maximization of R-squared and minimization of statistical errors.

Finally after the third phase, a prediction is realized to compute the probability of patients to develop diabetes. The methodology is also showed in the following figure:

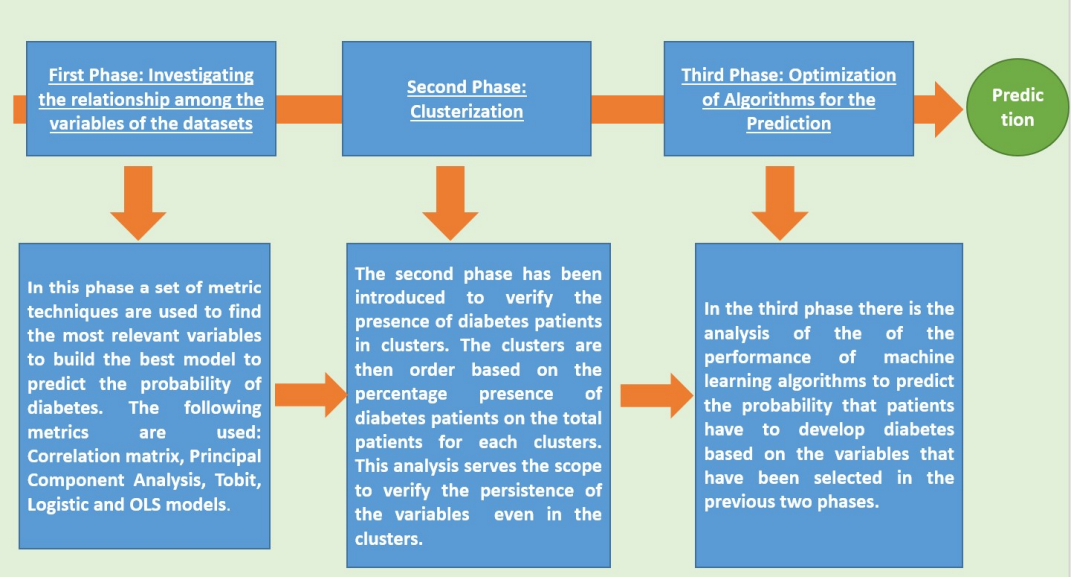

Figure 1. Methodology used to predict the probability that patients have to develop diabetes. The first phase is based on a set of metric analysis, the second phase is based on clusterization with fuzzy c-Means, the third phase contains a comparison among different machine learning algorithms to predict the probability of diabetes.

3. Literature Review

Diabetes and Pregnancies. Pre-pregnancy diabetes mellitus complicates approximately 0.3% of pregnancies in the US. While diabetes mellitus resulting from gestation complicates between 5% and 10% of pregnancies. The spread of type 2 diabetes among the pregnant population is the consequence of the spread of obesity among young US cohorts [2]. [3] define gestational diabetes mellitus as the most common complication of pregnancies with significant consequences for the health of mothers and new-borns. However, the authors emphasize that it is possible to prevent and promote greater health in women by reducing the risk of gestational diabetes mellitus by intervening in the correction of eating behaviours and in promoting sports activities of pregnant women in the periods of pre-conception, early pregnancy and interconception. [4] analyzed the relationship between gestational diabetes and external temperature at the time of delivery. The authors analyzed data from more than 24,000 pregnant women in an obstetric hospital in the UK. The results show that the consequences of gestational diabetes tend to worsen with increasing temperature. [5] check for a positive relationship between gestational diabetes and neonatal hypoglycaemia. [6] highlight the positive relationship between diabetes in pregnant women and the increase in infant mortality. [7] analyze the relationships between pregnant diabetic women and the Covid 19 pandemic. If pregnant women have normal weight or BMI overweight in the pre-twin pregnancy stages then there is a greater likelihood of gestational diabetes regardless of weight gain during pregnancy. [8] show a positive relationship between impaired right fetal heart function and gestational diabetes regardless of whether the pregnant woman is undergoing diabetes treatment or not. Pregnant women with obesity and gestational diabetes mellitus undergo changes in the function and size of the placenta [9]. [10] consider the risk of gestational diabetes mellitus in women undergoing assisted fertilization. [11] verify the presence of a positive relationship between pregnant women with diabetes and the presence of macrosomia in infants. The authors propose to use biomarkers to verify the presence of possible macrosomia in pregnant women with diabetes. [12] present the case of a diabetic patient who had two different pregnancies. In the first pregnancy, the pregnant woman with diabetes was treated with

punctures. Following the first pregnancy, the patient used the artificial pancreas. The artificial pancreas consists of an insulin pump, a continuous monitoring of blood glucose, and a control algorithm that regulates the administration of basal insulin. However, the use of artificial pancreas also requires the use of a certain interaction on the part of patients who have to interact with the open source algorithms. The use of the artificial pancreas appears to be very capable of improving the life of diabetics even if they express concerns about the secrecy of data and possible failures of the computer system. The woman who used the artificial pancreas in the second pregnancy said she had a better pregnancy than the first. The authors therefore propose to use the artificial pancreas for diabetic pregnant women to increase the quality of life of pregnant women. [13] analyze the effects of gestational diabetes in the case of single fetus and twin pregnancies. The authors verify that gestational diabetes mellitus is not associated with hypertension unlike single fetal pregnancies. However, the authors verify that there are also negative consequences in the case of gestational diabetes mellitus in the case of twin pregnancies such as accelerated growth of the fetus. [14] consider the determinants of GDM-gestational diabetes mellitus. Pregnant women are more likely to develop gestational diabetes mellitus in the presence of the following elements: overweight, obesity, advanced maternal age, family history of diabetes. Presence of gestational diabetes mellitus may increase the infant's risk of type 2 diabetes and macrosome. There is a positive relationship between gestational diabetes mellitus, body mass index and weight gain in pregnancy. Women who experience diabetes in pregnancy are also more likely to give birth to face the issue of the fetus or new-born Large for Gestational Age-LGA [15]. [16] highlight the negative consequences that the presence of diabetes mellitus can have on pregnant women and their offspring. This relationship is mainly connected to the growth of the Body Mass Index-BMI. [17] Found a positive relationship between the growth of the mean glucose value and the value of Large for Gestational Age-LGA and neonatal composite outcome-NCO. [18] consider the difference between Vaginal Delivery-VD and Caesarean Delivery-CD in a sample of Finnish women in childbirth characterized by gestational diabetes mellitus. The authors verify that the adverse consequences of childbirth tend to be greater in the case of using Caesarean Delivery-CD than with Vaginal Delivery-VD. [19] analyze stillbirth and infant mortality rates in connection with gestational diabetes mellitus. The authors collected data on stillbirths in connection with diabetic pregnancies in the period between 2006 and 2017 in Westmead Hospital. The data were obtained through the analysis of medical records. The authors analyzed the case of 37 women of which seven with type 1 diabetes, 11 with type 2 diabetes and 19 with gestational diabetes mellitus. The malformations that are connected with the presence of diabetes in pregnancy are mainly related to the cardiovascular and musculoskeletal dimensions. The authors conclude that stillbirth is still an unsolved problem in diabetic pregnancies. [20] identify a set of biomarkers that can be used to predict gestational diabetes mellitus. [21] verify the presence of a relationship between the presence of gestational diabetes mellitus and the onset of problems related to blood clotting for pregnant women. Through an empirical analysis [22] found a relationship between women with gestational diabetes mellitus and postpartum metabolic syndrome. [23] analyzed the relationship between gestational diabetes mellitus and the likelihood of developing diabetes in the later stages of a woman's life. The data show that the hazard rate for women who developed gestational diabetes mellitus is approximately 3.87 compared to women who did not develop gestational diabetes mellitus. That is, those who had diabetes during pregnancy are about 4 times more likely to develop type 2 diabetes in the 6-15 years following pregnancy. The authors also find a very long-term correspondence - up to 35 years - between those who had gestational diabetes mellitus and those who had type 2 diabetes following pregnancy.

Diabetes and Glucose. [24] verify the presence of a relationship between the interval time in the ambiance of continuous glucose monitoring and the value of diabetic retinopathy in patients with type

2 diabetes. [25] analyze the relationship between glucose-impaired metabolism and diabetes mellitus in the Basque region. The authors verify that 21.6% of the population had glucose-impaired metabolism. The factors that are associated with impaired glucose metabolism and diabetes are: male sex, abdominal obesity, high triglyceride levels, hypertension, family history of diabetes, age and cholesterol. [26] consider the presence of the relationship between glucose, weight and diabetes. The authors verify that weight reduction in the cohort of participants in the analysis has an absolutely positive effect in reducing the prediction of diabetes risk. Every kg of body weight lost reduces the risk of diabetes by 43%. The authors conclude that even small reductions in body weight can generate significant improvements in reducing the risk of diabetes. [27] refer to a drug capable of inhibiting the sodium-glucose cotransporter which is effective in reducing glycemic variability, regulating blood pressure and reducing body weight. [28] identify the presence of a threshold to distinguish between stable and unstable blood glucose in diabetic patients. In particular, the authors identify a measure that is $(\%CV = [(SD\ of\ Glucose) / (Mean\ Glucose) * 100)]$. The authors verify that when the %CV value exceeds 36% then the frequency of hypoglycemia increases especially in patients treated with insulin. [29] consider the impact on glucose metabolism of fucoidans useful in diabetes therapy. [30] offer a series of diabetes mellitus - defined as insufficient insulin production leading to a lack of proper blood glucose regulation. The authors distinguish among type 1 diabetes, type 2 diabetes and gestational diabetes. [31] analyze the case of the use of a digital diabetes level monitoring system called Smart Glucose Manager-SGM in Sri Lankan patients with diabetes. The analysis carried out with a group of users of the SGM system equal to 27 compared with a control group of 25 units, shows that the use of the Smart Glucose Manager-AGM has a positive impact in terms of improving the treatment of diabetes.

Diabetes and BMI. [32] show the limitations of the Body Mass Index-BMI in predicting Diabetes Remission. This predictive insufficiency of the Body Mass Index is associated with the existence of a non-linear relationship between muscle mass and the Body Mass Index. [33] find a non-linear relationship between BMI and HbA1c- average blood glucose sugar levels in a dataset of Danish children. [34] find a positive relationship between the increase in BMI and the spread of type 2 diabetes in Samoa by analysing data from a long historical series from 1978 to 2013 of 12,516 individuals. [35] analyze the impact of Roux-en-Y-RYBG gastric bypass in treating obese patients to reduce the impact of type 2 diabetes. The authors verify that patients with a reduced Body Mass Index-BMI have greater benefits from RYGB. [36] verify the existence of a positive relationship between Body Mass Index-BMI in a sample of 482,589 Chinese aged between 30 and 79 years between 2004 and 2008. The authors believe that the growth of adiposity in China indicates an increasing trend of diabetes in the population. [37] found that diabetic patients with a Body Mass Index-BMI between 27 and 34 had greater benefits from the combined use of bariatric surgery and intensive medical care compared to using intensive medical care alone. [38] verify a significant percentage of patients with diabetes remission for individuals with type 2 diabetes mellitus having a BMI <32.5 kg / m2 undergoing metabolic surgery in a sample of 112 Asians. [39] consider the relationship between patients with type 2 diabetes mellitus, the growth of the Body Mass Index-BMI and the likelihood of functional anomalies of the left atrium. [40] consider the existing relationship between the Body Mass Index-BMI and type 2 diabetes mellitus through the use of Mendelian randomization. [41] analyze the relationship between the Body Mass Index-BMI and the risk of type 2 diabetes in Chinese adults. [42] use machine learning techniques for the prevention of diabetes also through the use of the Body Mass Index-BMI. [43] analyze the role of obesity in determining diabetes in a multi-ethnic cohort of patients. The authors consider that type 2 diabetes is generally positively associated with breast cancer. However, the Body Mass Index for Latin women is not associated with the increased risk of breast cancer in patients with type 2 diabetes. [44] demonstrate the existence of

a positive relationship between the trend of the Body Mass Index-BMI and the risk of developing type diabetes in middle-aged women through the analysis of data of 12,302 women participating in the Australian Longitudinal Study on Women's Health -ALSWH between 1996 and 2016. [45] analyze the relationship between Body Mass Index-BMI and diabetes especially considering the limitations of the Body Mass Index as an indicator unable to distinguish between fat mass and lean mass. In fact, the authors also consider other indicators able to better represent the presence of fat mass and lean mass such as: waist circumference, the ratio between waist-height, predicted fat percentage, waist-hip ratio and body shape. [46] verify the presence of positive effects for diabetes patients with high basal BMI treated through the use of cinnamon.

Diabetes and Diabetes Pedigree Function. [47] find a positive correlation between Diabetes Pedigree Function, BMI and Glucose within a dataset aimed at diabetes detection. [48] use the variable Diabetes Pedigree Function to estimate diabetes through a set of machine learning algorithms measured by accuracy, precision and recall. [49] verify the existence of a directly proportional relationship between the variable Diabetes Pedigree Function and the value of Diabetes. [50] use diabetes pedigree function to estimate, together with other variables, the presence of diabetes through the use of the multilayer perceptron. [51] use the Pedigree Diabetes Function variable together with others with the Ensemble Machine Learning Techniques. [52] consider the positive role of diabetes pedigree function and weight in determining diabetes with machine learning algorithms. [53] uses the Pedigree Function Diabetes variable along with other variables to predict diabetes using fuzzy methods. [54] use the Pedigree Diabetes Function variable to predict diabetes with Artificial Neural Network. [55] use hypertuned machine learning techniques to predict diabetes mellitus with Pedigree Diabetics Function together with other variables. [56] verify the existence of a positive relationship between diabetes pedigree function and the presence of diabetes with a correlation index value equal to an amount of 0.17. [57] verify the presence of a relationship between diabetes and pedigree diabetes function through the Support Vector Machine. [58] it was estimated that in 2018 the number of diabetics in the world was equal to 451 million people and it is expected that this value in 2045 will reach an amount of 693 million people. [59] verify the presence of a positive relationship between Diabetes Pedigree Function and Diabetes through the use of a set of techniques and through the use of appropriate case studies. [60] demonstrate the presence of a positive relationship between diabetic pedigree function and diabetes prediction. [61] use a set of variables including Diabetes Pedigree Function for the prediction of diabetes mellitus through machine learning algorithms.

Finally, the literature suggests the presence of a positive relationship between aging and diabetes [62] and a positive relationship between diabetes and blood pressure [63].

Furthermore, in the analysis of the treatment of the main diseases it is necessary to consider the role of telemedicine [64] and the application [65] of medical devices [66] for monitoring and control [67]. Telemedicine is also relevant to offer home care that is essential in long and painful diseases as diabetes [68]. The application of AI algorithms can be useful in the treatment of diabetes [69]. Big data can offer a useful tool to improve the predictive performance of algorithms [70]. Finally, the use of AI can oriented the telemedicine and the health sector toward Industry 5.0 [71]. On a methodological point of view, the usage of augmented data can improve the application of predictive algorithm even in the management of healthcare datasets [72]. Similar considerations can also be realized for other disease that massively affect the global population such as hypertension [73]. Furthermore, the usage of machine learning can be effectively improve the ability to predict diabetes based on glycemic status of patients [74].

4. Data Description, Correlation Matrix and Principal Component Analysis

The statistical descriptions of the analyzed data[6] are presented below, that is:

- *Pregnancies:* has an average value of 3.7, a median value of 3, a minimum value of 0 and a maximum value of 17, a standard deviation value of 3.3061 units, a value of the coefficient of variation equal to an amount of 0.89269, an asymmetry value equal to 0.98163, a kurtosis value equal to 0.405;
- *Glucose:* has an average value of 121.18, a value of 117, a minimum value of 0, a maximum value of 199 units, a standard deviation value of 32.069, a value of the variation coefficient of 0 , 26463, an asymmetry value equal to 0.15869, a kurtosis value equal to 0.55597;
- *Blood Pressure:* average value equal to 69.14 units, a median value equal to 72, a value equal to 0, and a value equal to 122 units, the standard deviation value equal to 19,118 units, with a value of the coefficient of variation of 0.27751 units, an asymmetry value of -1.8531 units, a kurtosis value of 5.3122;
- *Skin Thickness:* has an average value of 20.93 units, a median value of 23, a minimum value of 0 and a maximum value of 110 units, a standard deviation value of 16,103 units, with a value the variation coefficient equal to 0.7692, an asymmetry value equal to 0.207, a kurtosis value equal to 0.15219;
- *BMI-Body Mass Index:* with an average value of 32.19, a median value of 32.3, a minimum value of 0, a maximum value of 80.6, with a standard deviation value of 8.149, a value of the variation coefficient equal to 0.25316, asymmetry equal to a value of -0.090387, kurtosis with a value equal to 4.1184;
- *Diabetes Pedigree Function:* with an average value of 0.47, a median value of 0.376, a minimum value of 0.078 units and a maximum of 2.42 units, standard deviation with a value of 0.32355, with an evaluation coefficient value equal to 0.68705, an asymmetry value equal to 1.8106, a kurtosis value equal to 4.9913;
- *Age*: with an average value of 33.09, a median value of 29, a minimum value of 21, a maximum value of 81, standard deviation with a value of 11,786 units, with a value of the coefficient of variation equal to an amount of 0.35619, an asymmetry value equal to an amount of 1.1804, a kurtosis value equal to 0.82132**.**

[6] Data are collected from Kaggle: https://www.kaggle.com/datasets/johndasilva/diabetes

Descriptive Statistics								
Variables	**Average**	**Median**	**Minimum**	**Maximum**	**Standard Deviation**	**Coefficient of variation**	**Asymmetry**	**Kurtosis**
Pregnancies	3,7035	3	0	17	3,3061	0,89269	0,98163	0,40585
Glucose	121,18	117	0	199	32,069	0,26463	0,15869	0,55597
BloodPressure	69,145	72	0	122	19,188	0,27751	-1,8531	5,3122
BMI	32,193	32,3	0	80,6	8,1499	0,25316	-0,090387	4,1184
Diabetes Pedigree Function	0,47093	0,376	0,078	2,42	0,32355	0,68705	1,8106	4,9913
Age	33,09	29	21	81	11,786	0,35619	1,1804	0,82132
Diabetes	0,342	0	0	1	0,4745	1,3874	0,66613	-1,5563

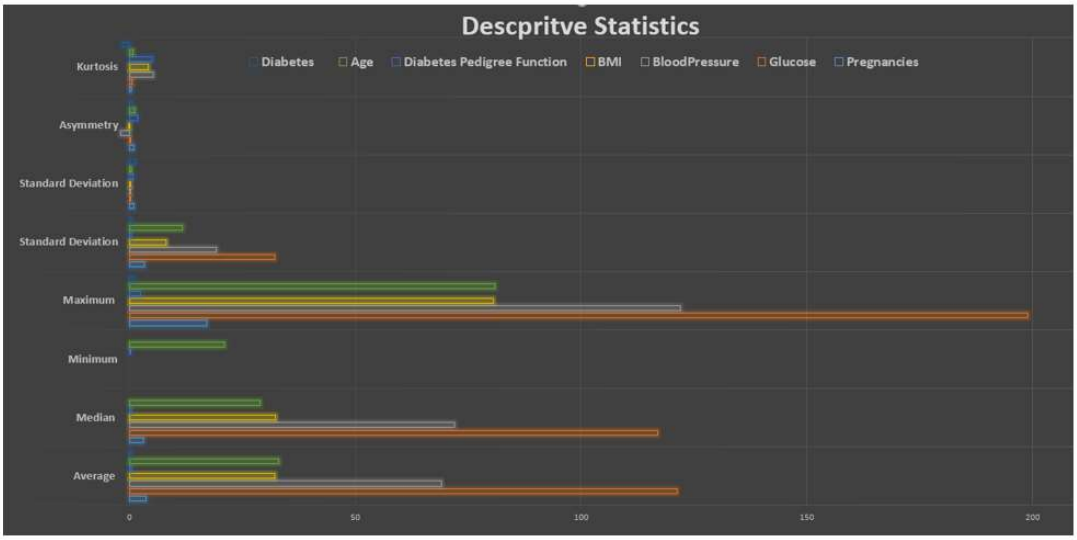

Figure 2. Descriptive Statistics of the dataset. In this figure the main characteristics of the variables of the dataset are presented in order to introduce the metric analysis oriented to identify the variables of the model.

A correlation matrix is then created to verify the relationships of the model variables. Diabetes is therefore associated with the following variables:

- *Blood Pressure* with a value of 0.0760;
- *Diabetes Pedigree Function* with a value of 0.1555;
- *Pregnancies* with a value of 0.2244;
- *Age* with a value of 0.2365;
- *BMI-Body Mass Index* with a value equal to 0.2767;
- *Glucose* with a value of 0.4584.

Diabetes Correlations Matrix. Frankfurt Hospital

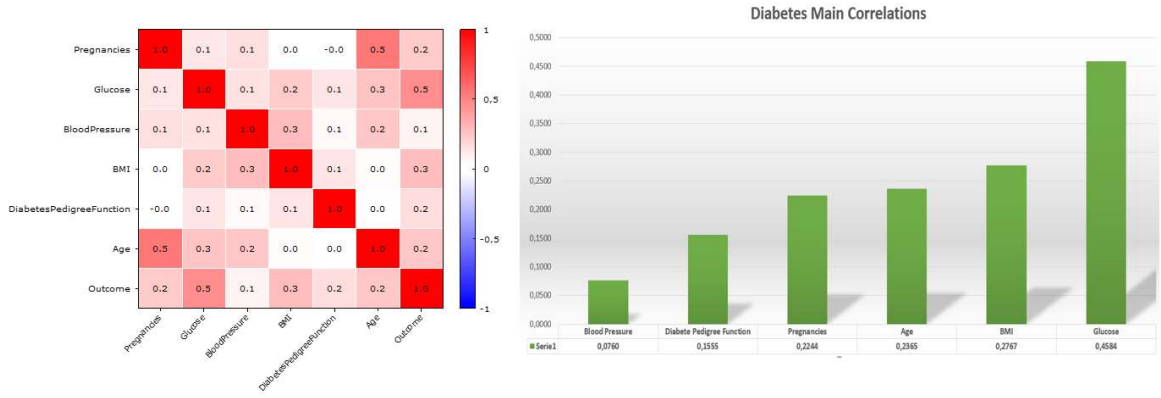

Figure 3. Correlation Matrix Diabetes. The correlation matrix has the ability to show the presence of specific relationships that can support the building process of an optimal model to predict the probability of diabetes in the patients. As shows in the figure, it is clear the presence of a relationship between glucose and diabetes, and the presence of a relationship between diabetes and BMI. Those two relationships have the higher level of correlation index.

A Principal Component Analysis-PCA is carried out below to verify the presence of a relationship between the variables of the observed model:

- *PCA1*: show the presence of a positive relationship between "*Diabetes*", "*Age*" and "*Glucose*". In particular, the "*Diabetes*" value is 0.465, the "*Age*" value is 0.457 and the "*Glucose*" value is 0.44. It follows that "*Age*", "*Diabetes*" and "*Glucose*" are positively associated or grow together;
- *PCA2*: the most significant variables are "*BMI-Body Mass Index*" with a value of 0.463, followed by "*Age*" with a value of -0.467, and "*Pregnancies*" -0.546 units. It therefore follows that "*BMI-Body Mass Index*" exists on the one hand and "Age" and Pregnancies on the other grow inversely proportional;
- *PCA3*: the most significant variables are "*Diabetes*" with a value of 0.378, "*BMI-Body Mass Index*" with a value of -0.195, and "*Blood Pressure*" with a value of -0.751 units. It therefore follows that "*Diabetes*" on the one hand and "*BMI-Body Mass Index*" and "*Blood Pressure*" on the other have an inversely proportional trend;
- *PCA4*: the most significant variables are "*Diabetes Pedigree Function*" with a value equal to 0.879, and "*Diabetes*" with a value equal to -0.227 and "*Glucose*" with an amount equal to -0.292. It follows that there is an inverse relationship between a "*Diabetes Pedigree Function*" on the one hand and "*Diabetes*" and "*Glucose*";
- *PCA5:* the most significant variables are "*BMI-Body Mass Index*" with a value equal to 0.582, "*Blood Pressure*" with a value equal to -0.422, and "*Glucose*" equal to -0.55. Therefore it follows that there is an inverse relationship between "*Body Mass Index-BMI*" on the one hand and "*Blood Pressure*" and "*Glucose*" on the other;
- *PCA6:* the most significant variables are "*Glucose*" with a value equal to 0.387, and "*Blood Pressur*e" with a value equal to -0.375, and "*Diabetes*" with a value equal to -0.669. The result

therefore is the presence of an inverse relationship between "*Glucose*" on the one hand and "*Diabetes*" and "*Blood Pressure*" on the other;

- *PCA7:* the most significant variables are "*Age*" with a value of 0.656, "*Glucose*" with a value of -0.34 and "*Pregnancies*" with a value of -0.619. It therefore follows that there is a negative relationship between "*Age*" on the one hand and "*Glucose*" and "*Pregnancies*" on the other.

In this section the description of the dataset, the correlation matrix and the PCA have showed the presence of certain relationships among the variables that can be used to estimate the value of diabetes that are analyzed deeply in the following paragraph.

5. The Econometric Model to Estimate the Value of Diabetes

We estimated the presence of diabetes through a set of econometric models, namely: Tobit, OLS-Ordinary Least Squares, Logit Multinomial, and WLS Corrected for Heteroskedasticity, Probit, and Logit. The use of the Multinomial Logit, Probit, Logit and Tobit models was necessary since the estimated variable is dichotomous and assumes a value of 0 or alternatively 1. Specifically we have estimated the following model:

$$\boldsymbol{Diabetes_i = a_1 + b_1(Pregnancies)_i + b_2(Glucose)_i + b_3(BloodPressure)_i + b_4(BMI)_i + b_5(DiabetesPedigreeFunction)_i + b_6(Age)_i}$$

$$\boldsymbol{i = 2000}$$

In particular, it appears that diabetes is positively associated with:

- *Pregnancies*: There is a positive relationship between diabetes and pregnancies. This value is 0.054 in the case of the Tobit model, 0.022 in the case of the OLS, 0.1268 in the case of the Multinomial Logit, 0.02075 in the case of the WLS corrected for heteroskedasticity, 0.0744 in the case of the Probit, 0 , 1268 in the case of the Logit. On average, pregnancies is related to diabetes with a value of 0.070.
- *Glucose*: assumes a value equal to 0.014 in the case of the Tobit model, 0.0055 in the case of the OLS, 0.032116 in the case of the Multinomial Logit, 0.001 in the case of the WLS corrected for heteroskedasticity, 0.0179 in the case of Probit, 0.032 in the case of Logit. On average, the impact of glucose in determining diabetes is equal to an amount of 0.017 uintà.
- *BMI-Body Mass Index:* is associated with Diabetes with a value equal to 0.0354 units in the case of Tobit, 0.011 in the case of the OLS, 0.075 in the case of the multinomial Logit, 0.0103 in the case of the WLS corrected for the heteroskedasticity, 0.0425 in the case of the Probit model, 0.0751 in the case of the Logit model. On average, the relationship between BMI and diabetes is equal to an amount of 0.041 units.
- *Diabetes Pedigree Function*: it is positively associated with diabetes with a value of 0.3189 in the case of the Tobit model, equal to an amount of 0.13245 in the case of the OLS model, equal to an amount of 0.8277 in the case of Logit multinomial, 0.14601 in the case of WLS corrected for heteroskedasticity, 0.437981 in the case of the Probit model, 0.8277 in the case of the Logit model. On average, for the models analyzed, the value is equal to an amount of 0.448497.
- *Age*: has a value of 0.0100884 in the case of the Tobit model, equal to 0.00262656 in the case of the OLS model, equal to 0.151539 in the case of the multinomial Logit, equal to a value of 0.00764 in the case of WLS corrected for heteroskedasticity, equal to 0.00973 in the case of the Probit, and equal to a value of 0.0151 in the case of the Logit model. On average, this value has an impact of 0.01 in the determination of diabetes within the considered model.

Furthermore, diabetes is negatively associated with the following variable:

- *Blood Pressure*: assumes a value equal to -0.0042 in the case of the Tobit model, equal to an amount of -0.001 in the case of the OLS model, equal to an amount of -0.0099 in the case of the multinomial Logit model, equal to a value of -0.00213 in the case of the WLS model corrected for heteroskedasticity, equal to an amount of -0.005467 in the case of the Probit model, -0.00996126 in the case of the Logit model, and on average equal to an amount of -0.0047.

In this analysis a series of metric model has been used to estimate the value of diabetes. Different models have been used in order to have confirmations about the persistence of investigates relationship. Specifically we have used two different kinds of metric models i.e. binary models and non binary models. Binary models are: Tobit, Probit, Logit, Multinomial Logit. Non binary models are: OLS and WLS corrected for heteroskedasticity. The use a differentiated set of models is necessary to investigate the persistence of the relationships either in the respect to the functional framework of the dataset that is based on a binary variable, either in the context of noon-binary models. The confrontation among those models show that the investigated variables are persisten and in a certain sense model-indifferent even if the most appropriate models to investigate the data are the binary models. Furthermore, it is possible to observe that the results are substantially consistent with the literature with the exception of the negative relationship between diabetes and blood pressure. In effect, the negative relationship between diabetes and blood pressure that the confrontation among binary and non-binary model suggests, is in contrast with the main assumptions of the literature. This contrast between the literature review and the empirical analysis show a case sensitivity in the analyzed data.

Estimation of Diabetes								
Variables		**Const**	**Pregnancies**	**Glucose**	**Blood Pressure**	**BMI**	**Diabetes Pedigree Function**	**Age**
Tobit Model	***Coefficient***	-3,60329	0,0547208	0,0141852	-0,0042836	0,0354085	0,318987	0,0100884
	p-Value	***	***	***	***	***	***	***
OLS	***Coefficient***	-0,80539	0,0220046	0,00559678	-0,0018368	0,0113498	0,13245	0,00262656
	p-Value	***	***	***	***	***	***	***
Logit Multinomiale	***Coefficient***	-7,87008	0,126861	0,032116	-0,00996126	0,0751571	0,827777	0,0151539
	p-Value	***	***	***	***	***	***	***
WLS corrected for heteroskedasticity	***Coefficient***	-0,39678	0,02075	0,001	-0,00213	0,0103	0,14601	0,00764
	p-Value	***	***	***	***	***	***	***
Probit	***Coefficient***	-4,49837	0,0744666	0,0179265	-0,005467	0,0425482	0,437981	0,00973062
	p-Value	***	***	***	***	***	***	***
Logit	***Coefficient***	−7,87008	0,126861	0,032116	−0,00996126	0,0751571	0,827777	0,0151539
	p-Value	***	***	***	***	***	***	***
Mean		-3,43478	0,070944	0,017156747	-0,004735732	0,04165345	0,448497	0,010065563

Figura 4. Synthesis of the metric results for the estimation of diabetes. The model show a confrontation between binary and non-binary models. The combined use of binary and non-binary models is proposed to verify the cross model persistency of the proposed model. Results show that the investigation of diabetes is model-indifferent even if the binary model are the most appropriate to estimate the level of diabetes.

6. Clusterization with Fuzzy c-Means optimized with Elbow Method

A clustering analysis is then determined to verify the characteristics of the production of the individual clusters analyzed. The use of the Elbow method allows you to graphically verify the optimal number of clusters considering the relationship between the "*Sum of Squares Distances from Centroids*" and the number of clusters k. Through this analysis, it results that the optimal number of

clusters is equal to 3. Therefore, using the fuzzy c-Means clustering model by imposing a value of k = 3 it was possible to calculate the characteristics of the single clusters. It should be considered that the data that is entered into the clustering scheme are subjected to partitioning, that is: 70% of the data is used to train the algorithm while the remaining 30% was used for actual clustering. Therefore, the following characteristics of the clusters derive, namely:

- *Cluster 0 = C0*: has a median value of "*Pregnancies*" equal to 3, a median value of "*Glucose*" equal to 140.00, a "*Blood Pressure*" value equal to 76.00, a median value of "*Body Mass Index-BMI*" equal to 34.30, a median value of "*Diabetes Pedigree Function*" equal to 0.36, a median value of "*Age*" equal to 29.00. In this clusters there is a number of diabetics equal to 121.00 with a percentage of diabetics on the total records equal to 43.53%;
- *Cluster 1 = C1*: has a median "*Pregnancies*" value of 4.00 units, a median "*Glucose*" value of 116.50 units, a "*Blood Pressure*" value of 64.00 units, a median value of "*BMI-Body Mass Index*" equal to 27.70 units, a value of "*Diabetes Pedigree Function*" equal to a value of 0.33, a median value of "*Age*" equal to 38.50 units. In this cluster there are 44 diabetics, out of a total number of records equal to 120.00 units and a percentage of diabetics equal to 36.67%;
- *Cluster 2 = C2:* with a median number of "*Pregnancies*" equal to 2,000, a median value of "*Glucose*" equal to 91.50, a median value of "*Blood Pressure*" equal to 70.00, a median value of "*Age*" equal to 26.00 units. The total number of diabetics present within the cluster considered is equal to 19.00 units, the total number of records equal to 202 with an incidence of diabetics equal to an amount of 9.41%.

From the point of view of the percentage presence of diabetics in the clusters, it is possible to identify the following order, namely C0 = 43.53%> C1 = 36.67%> C2 = 9.41%. It is therefore possible to calculate the value of the individual variables of Cluster 0-C0, or the clusters with a greater number of diabetics, compared to the values of Cluster 1-C1 and Cluster 2-C2. The value of "*Pregnancies*" of Cluster 0-C0 is equal to a value of 75% with respect to the value of Cluster 1-C1 and equal to a value of 150% of the value of Cluster 2-C2. The "*Glucose*" value of Cluster 0-C0 is equal to an amount of 120.17% of Cluster 1-C1, and equal to a value of 153.00% of the corresponding value of Cluster 2-C2. The "*Blood Pressure*" value of Cluster 0-C0 is equal to an amount of 119% of the value of Cluster 1-C1, and equal to an amount of 109% of Cluster 2-C2. "*BMI-Body Mass Index*" of Cluster 0-C0 is equal to an amount of 123.82% of the value of Cluster 1-C1, while the value of cluster 2-C2 of the "*Body Mass Index*" is equal to an amount of 116.271%. The value of "Diabetes Pedigree Function" of Cluster 0-C0, is equal to an amount of 109.84% of Cluster 1-C1, and a value of Cluster

2-C2, equal to a value of 92.60%. The Age value of Cluster 0-C0 is equal to 75.32% of Cluster 1-C1 and equal to a value of 111.53% of Cluster 2-C2.

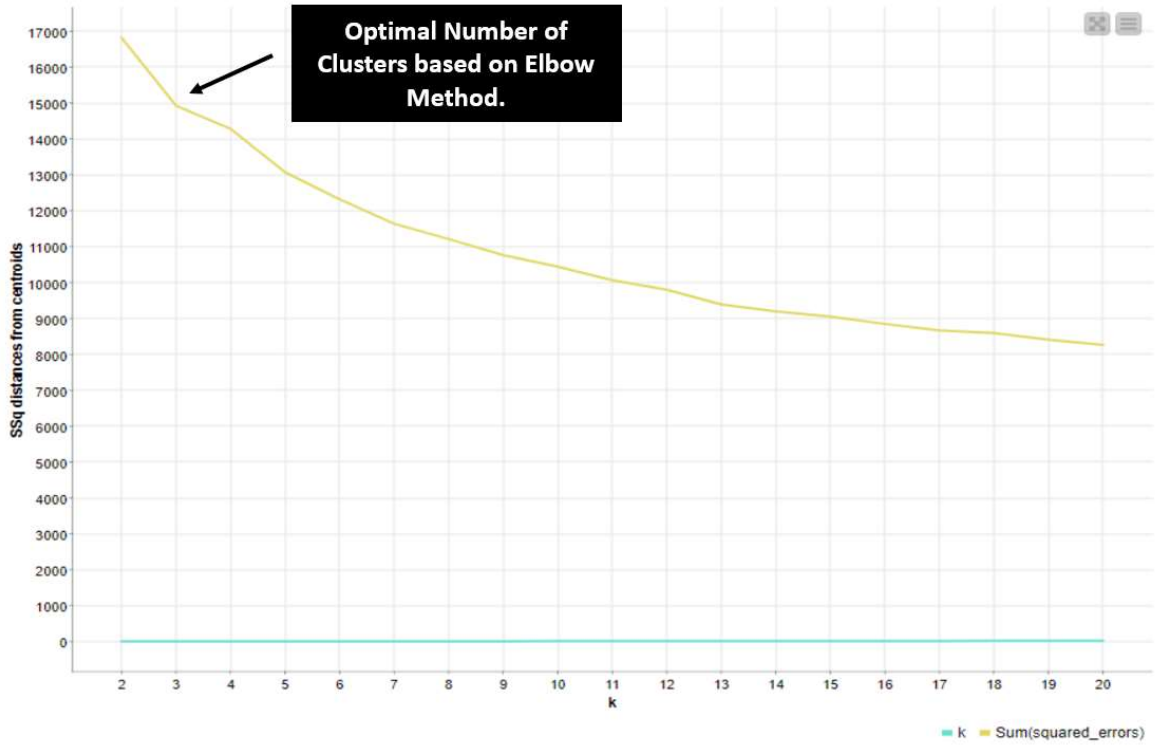

Figure 5. The use of Elbow Method is an alternative in respect to Silhouette coefficient to estimate the optimal number of cluster in a dataset. In this case the optimal number of clusters is equal to 3.

7. Machine Learning and Prediction

Machine learning algorithms for predicting the value of diabetes are presented below. In particular, four different statistical indicators were used, namely: R-squared, Mean Absolute Error, Mean Squared Error, Root Mean Squared Error. Specifically, the algorithms were trained with 70% of the available data while the remaining 30% was used for the actual prediction. The results show the following ranking of the algorithms by predictive capacity, that is:

- Tree Ensemble Regression with a payoff value of 5;
- Random Forest Algorithm with a payoff value of 9;
- Gradient Boosted Trees and Simple Regression Tree with a payoff value of 13;
- ANN-Artificial Neural Network with a payoff value of 21;
- Polynomial Regression with a payoff value of 25;
- Linear Regression with a payoff value of 29;
- PNN-Probabilistic Neural Network with a payoff value of 30.

Ranking of Algorithms by Predictive Performance						
Rank	**Algorithms**	***R^2***	***Mean absolute error***	***Mean squared error***	***Root mean squared error***	***Sum***
1	*Tree Ensemble Regression*	1	2	1	1	5
2	*Random Forest*	2	3	2	2	9
3	*Gradient Boosted Trees*	3	4	3	3	13
3	*Simple Regression Tree*	4	1	4	4	13
4	*ANN-Artificial Neural Network*	5	6	5	5	21
5	*Polynomial Regression*	6	7	6	6	25
6	*Linear Regression*	7	8	7	7	29
7	*PNN-Probabilistic Neural Network*	8	5	8	9	30

Figure 6. Ranking of Algorithms by predictive performance. Each column represents a ranking and each algorithm has a position in that ranking. The rank that algorithms have in each rank is summed up in a final ranking. The main statistical measure analysed are R-Squared, Mean Absolute Error, Mean Squared Error, Root Mean Squared Error.

Statistical Results of Machine Learning Algorithms for the Prediction of Diabetes				
Statistical Results	***R^2***	***Mean absolute error***	***Mean squared error***	***Root mean squared error***
ANN-Artificial Neural Network	0,3766206	0,283943479	0,136019654	0,368808424
PNN-Probabilistic Neural Network	0,01831502	0,223333333	0,223333333	0,472581563
Simple Regression Tree	0,49494949	0,111666667	0,111666667	0,334165628
Gradient Boosted Trees	0,65505237	0,193849418	0,079475935	0,281914765
Random Forest	0,71253568	0,17949583	0,06272392	0,25044743
Linear Regression	0,21998106	0,34073427	0,17290203	0,41581490
Polynomial Regression	0,32799341	0,32758129	0,15859169	0,39823572
Tree Ensemble Regression	0,72996779	0,17749881	0,062586716	0,250173373

Figura 7. Statistical Results of Machine Learning Algorithms for the Prediction of Diabetes. The figure shows the number that are the result of machine learning algorithm.

Therefore, using the Tree Ensemble Regression algorithm, which is the best performing algorithm, it was possible to realize the following prediction, namely:

$$f(\boldsymbol{Diabetes}) = \begin{cases} \boldsymbol{DiabetesRisk = 0}, & \boldsymbol{Diabetes_{Prediction} < 0,5} \\ \boldsymbol{DiabetesRisk = 0,5}, & \boldsymbol{0,5 < Diabetes_{Prediction} < 0,75} \\ \boldsymbol{DiabetesRisk = 1}, & \boldsymbol{Diabetes_{Prediction} > 0,75} \end{cases}$$

Prediction of the Value of Diabetes		
Diabetes Value	**Patients**	**Description**
1<x<0,75	*137*	*The value of diabetes appears to be between a value of 1 and a value of 0.75 units. In this case, the risk of diabetes is considered maximum, i.e. equal to 1. In the analyzed dataset there are therefore 137 people for whom a diabetes value of 1 was predicted.*
0,75<x<0,50	*78*	*In the analyzed dataset, the predicted value of diabetes is between 0.75 and 0.50 for 78 people. It follows that for this part of the population analyzed there is a risk of incipient diabetes. That is, these people could easily develop diabetes.*
0,5<x<0,0	*385*	*385 people have a predicted diabetes risk value between 0.0 and 0.5 or essentially equal to 0.*

Figure 8. Prediction of the Value of Diabetes. The table synthetizes the results of the prediction considering the probability to develop diabetes based on the best performing algorithms.

8. Conclusions

The presented article show a complex methodology to selected the best variables to predict the probability of a patient to develop diabetes among a set of variables. Data are collected from 2000 patients from a Hospital in Frankfurt. The dataset is public and has been published on Kaggle. The methodological approach used has been based on three phases: the first phase is based on a set of metric test to select the most relevant variables to use in the prediction, the second phase investigates the persistency of the relationships among the variables in a set of clusters optimized with the Elbow method, the third phase presents a confrontation among eight different machine learning algorithms to estimate the probability of patients to develop diabetes. Even if the applied methodology shows some elements of originality, the results are consistent with the literature with the exception of the negative relationship between blood pressure and diabetes that can be explained as a consequence of the specifically characteristics of the dataset.

9. Declarations

Author Contributions. Conceptualization, A.L., N.M., G.C.,V.G., and A.M.; methodology, A.L.; software, G.C.; validation, N.M., G.C., V.G.and A.M.; formal analysis, A.L.; investigation, N.M.; resources, G.C. and V.G.; data curation, A.M.; writing—original draft preparation, A.L.; writing—review and editing, N.M.; visualization, G.C. and V.G.; supervision, A.M.; project administration, A.L. All authors have read and agreed to the published version of the manuscript.

Data Availability Statement. The data presented in this study are available on request from the corresponding author.

Funding. The authors received no financial support for the research, authorship, and/or publication of this article.

Declaration of Competing Interest. The authors declare that there is no conflict of interests regarding the publication of this manuscript. In addition, the ethical issues, including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, and redundancies have been completely observed by the authors.

10. Figure Index

References

[1] H. Hall, D. Perelman, A. Breschi, P. Limcaoco, R. Kellogg, T. McLaughlin and M. Snyder, "Glucotypes reveal new patterns of glucose dysregulation," *PLoS biology,* vol. 7, no. e2005143, p. 16, 2018.

[2] D. R. Coustan, "Diabetes in pregnancy," *Clinical Maternal-Fetal Medicine Online ,* pp. 16-1, 2021.

[3] D. A. Schoenaker, S. De Jersey, J. Willcox, M. E. Francois and S. Wilkinson, "Prevention of gestational diabetes: the role of dietary intake, physical activity, and weight before, during, and between pregnancies," *Seminars in reproductive medicine,* vol. 6, no. 38, pp. 352-365, 2020.

[4] C. L. Meek, B. Devoy, D. Simmons, C. J. Patient, A. R. Aiken, H. R. Murphy and C. E. Aiken, "Seasonal variations in incidence and maternal–fetal outcomes of gestational diabetes," *Diabetic Medicine,* vol. 37, no. 4, pp. 674-680, 2020.

[5] A. C. Sheehan, M. P. Umstad, S. Cole and T. J. Cade, "Does gestational diabetes cause additional risk in twin pregnancy?," *Twin Research and Human Genetics,* vol. 22, no. 1, pp. 62-69, 2019.

[6] A. P. Sunjaya and A. F. Sunjaya, "Diabetes in pregnancy and infant mortality: Link with glycemic control," *Diabetes & Metabolic Syndrome: Clinical Research & Reviews,* vol. 12, no. 6, pp. 1031-1037., 2018.

[7] C. Eberle, T. James-Todd and S. Stichling, "SARS-CoV-2 in diabetic pregnancies: a systematic scoping review," *BMC pregnancy and childbirth,* vol. 21, no. 1, pp. 1-10, 2021.

[8] F. D'Ambrosi, G. Rossi, C. M. Soldavini, I. F. Carbone, G. E. Cetera, N. Cesano and E. Ferrazzi, "Evaluation of fetal cardiac function in pregnancies with well-controlled gestational diabetes," *Archives of Gynecology and Obstetrics,* vol. 304, no. 2, 2021.

[9] K. Tanaka, K. Yamada, M. Matsushima, T. Izawa, S. Furukawa, Y. Kobayashi and M. Iwashita, " Increased maternal insulin resistance promotes placental growth and decreases placental efficiency in pregnancies with obesity and gestational diabetes," *Journal of Obstetrics and Gynaecology Research,* vol. 44, no. 1, pp. 74-80, 2018.

[10] A. Kouhkan, M. E. Khamseh, A. Moini, R. Pirjani, A. E. Valojerdi, A. Arabipoor and H. R. Baradaran, "Predictive factors of gestational diabetes in pregnancies following assisted reproductive technology: a nested case–control study," *Archives of gynecology and obstetrics, ,* vol. 298, no. 1, pp. 199-206, 2018.

[11] S. Nahavandi, J. M. Seah, A. Shub, C. Houlihan and E. I. Ekinci, "Biomarkers for macrosomia prediction in pregnancies affected by diabetes," *Frontiers in endocrinology,* vol. 9, no. 407, 2018.

[12] I. Schütz-Fuhrmann, A. K. Schütz, M. Eichner and J. K. Mader, "Two subsequent pregnancies in a woman with type 1 diabetes: artificial pancreas was a gamechanger," *Journal of Diabetes Science and Technology,* vol. 14, no. 5, pp. 972-973, 2020.

[13] L. Hiersch, H. Berger, R. Okby, J. G. Ray, M. Geary, S. D. McDonald and N. Melamed, "Gestational diabetes mellitus is associated with adverse outcomes in twin pregnancies," *American journal of obstetrics and gynecology,* vol. 1, no. 102-e1, p. 220, 2019.

[14] J. F. Plows, J. L. Stanley, P. N. Baker, C. M. Reynolds and M. H. Vickers, "The pathophysiology of gestational diabetes mellitus," *International journal of molecular sciences,* vol. 11, no. 3342, p. 19, 2018.

[15] F. Weschenfelder, F. Hein, T. Lehmann, E. Schleußner and T. Groten, "Contributing factors to perinatal outcome in pregnancies with gestational diabetes—what matters most? A retrospective analysis.," *Journal of Clinical Medicine,* vol. 2, no. 348, p. 10, 2021.

[16] C. M. Reynolds, E. G. O'Malley, B. Egan, S. R. Sheehan and M. J. Turner, "Maternal weight trajectories in successive pregnancies and their association with gestational diabetes mellitus," *Diabetes Care,* vol. 43, no. 3, pp. e33-e34, 2020.

[17] K. Kristensen, L. E. Ögge, V. Sengpiel, K. Kjölhede, A. Dotevall, A. Elfvin and K. Berntorp, "Continuous glucose monitoring in pregnant women with type 1 diabetes: an observational cohort study of 186 pregnancies," *Diabetologia,* vol. 7, no. 1143, p. 62, 2019.

[18] H. Kruit, S. Mertsalmi and L. Rahkonen, "Planned vaginal and planned cesarean delivery outcomes in pregnancies complicated with pregestational type 1 diabetes–A three-year academic tertiary hospital cohort study," *BMC Pregnancy and Childbirth,* vol. 22, no. 1, 2022.

[19] M. Wang, N. Athayde, S. Padmanabhan and N. W. Cheung, "Causes of stillbirths in diabetic and gestational diabetes pregnancies at a NSW tertiary referral hospital," *Australian and New Zealand Journal of Obstetrics and Gynaecology,* vol. 59, no. 4, pp. 561-566, 2019.

[20] P. Wu, W. E. Farrell, K. E. Haworth, R. D. Emes, M. O. Kitchen, J. R. Glossop and A. A. Fryer, "Maternal genome-wide DNA methylation profiling in gestational diabetes shows distinctive disease-associated changes relative to matched healthy pregnancies," *Epigenetics,* vol. 13, no. 2, pp. 122-128, 2018.

[21] J. Teliga-Czajkowska, J. Sienko, J. Zareba-Szczudlik, A. Malinowska-Polubiec, E. Romejko-Wolniewicz and K. Czajkowski, "Influence of glycemic control on coagulation and lipid metabolism in pregnancies

complicated by pregestational and gestational diabetes mellitus," *Advances in Biomedicine ,* pp. 81-88, 2019.

[22] Y. Shen, W. Li, J. Leng, S. Zhang, H. L. W. Liu and G. Hu, "High risk of metabolic syndrome after delivery in pregnancies complicated by gestational diabetes," *Diabetes research and clinical practice,* no. 150, pp. 219-226, 2019.

[23] M. V. Diaz-Santana, K. M. O'Brien, Y. M. M. Park, D. P. Sandler and C. R. Weinberg, "Persistence of risk for type 2 diabetes after gestational diabetes mellitus," *Diabetes Care,* vol. 45, no. 4, pp. 864-870, 2022.

[24] J. Lu, X. Ma, J. Zhou, L. Zhang, Y. Mo, L. Ying and W. Jia, "Association of time in range, as assessed by continuous glucose monitoring, with diabetic retinopathy in type 2 diabetes," *Diabetes Care,* vol. 41, no. 11, pp. 2370-2376, 2018.

[25] A. Aguayo, I. Urrutia, T. González-Frutos, R. Martínez, L. Martínez-Indart, L. Castaño and A. Garrido, "Prevalence of diabetes mellitus and impaired glucose metabolism in the adult population of the Basque Country, Spain," *Diabetic Medicine,* 2017.

[26] K. I. Galaviz, M. B. Weber, A. Straus, J. S. Haw, K. V. Narayan and M. K. Ali, "Global diabetes prevention interventions: a systematic review and network meta-analysis of the real-world impact on incidence, weight, and glucose.," *Diabetes Care,* vol. 41, no. 7, pp. 1526-1534, 2018.

[27] T. Danne, S. Garg, A. L. Peters, J. B. Buse, C. P. J. H. Mathieu and M. Phillip, "International consensus on risk management of diabetic ketoacidosis in patients with type 1 diabetes treated with sodium–glucose cotransporter (SGLT) inhibitors," *Diabetes care,* vol. 42, no. 6, pp. 1147-1154, 2019.

[28] L. Monnier, C. Colette, A. Wojtusciszyn, S. Dejager, E. Renard, N. Molinari and D. R. Owens, "Toward defining the threshold between low and high glucose variability in diabetes," *Diabetes care,* vol. 40, no. 7, pp. 832-838, 2017.

[29] B. Mabate, C. D. Daub, S. Malgas, A. L. Edkins and B. I. Pletschke, "Fucoidan structure and Its impact on glucose metabolism: Implications for diabetes and cancer therapy," *Marine Drugs,* vol. 1, no. 30, p. 19, 2021.

[30] K. Turksoy, E. Littlejohn and A. Cinar, "Multimodule, multivariable artificial pancreas for patients with type 1 diabetes: regulating glucose concentration under challenging conditions," *IEEE Control Systems Magazine,* vol. 38, no. 1, pp. 105-124, 2018.

[31] K. C. Gunawardena, R. Jackson, I. Robinett, L. Dhaniska, S. Jayamanne, S. Kalpani and D. Muthukuda, "The influence of the smart glucose manager mobile application on diabetes management," *Journal of diabetes science and technology,* vol. 13, no. 1, pp. 75-81, 019.

[32] S. Li, H. Yu, P. Zhang, Y. Tu, Y. Xiao, D. Yang and W. Jia, "The Nonlinear Relationship Between Psoas Cross-sectional Area and BMI: A New Observation and Its Insights Into Diabetes Remission After Roux-en-Y Gastric Bypass," *Diabetes Care,* vol. 44, no. 12, pp. 2783-2786, 2021.

[33] E. H. Ibfelt, D. Vistisen, P. Falberg Rønn, S. Pørksen, M. Madsen, B. Kremke and J. Svensson, "Association between glycaemic outcome and BMI in Danish children with type 1 diabetes in 2000–2018: a nationwide population-based study," *Diabetic Medicine,* vol. 3, no. e14401, p. 38, 2021.

[34] S. Lin, T. Naseri, C. Linhart, S. Morrell, R. Taylor, S. T. McGarvey and P. Zimmet, "Trends in diabetes and obesity in Samoa over 35 years, 1978–2013," *Diabetic Medicine,* vol. 34, no. 5, pp. 654-661, 2017.

[35] G. Ji, W. Li, P. Li, H. Tang, Z. Yu, X. Sun and S. Zhu, "Effect of Roux-en-Y gastric bypass for patients with type 2 diabetes mellitus and a BMI< 32.5 kg/m2: a 6-year study in Chinese patients," *Obesity Surgery,* vol. 30, no. 7, pp. 2631-2636, 2020.

[36] F. Bragg, K. Tang, Y. Guo, A. Iona, H. Du and M. V. Holmes, "Associations of general and central adiposity with incident diabetes in Chinese men and women," *Diabetes care,* vol. 41, no. 3, pp. 494-502, 2018.

[37] P. R. Schauer, D. L. Bhatt, J. P. Kirwan, K. Wolski, A. Aminian, S. A. Brethauer and S. R. Kashyap, "Bariatric surgery versus intensive medical therapy for diabetes—5-year outcomes," *N Engl J Med,* vol. 376, pp. 641-651, 2017.

[38] Z. Yu, W. Li, X. Sun, H. Tang, P. Li, G. Ji and S. Zhu, "Predictors of Type 2 Diabetes Mellitus Remission After Metabolic Surgery in Asian Patients with a BMI< 32.5 kg/m2," *Obesity Surgery,* vol. 31, no. 9, pp. 4125-4133, 2021.

[39] M. T. Shen, Y. K. Guo, X. Liu, Y. Ren, L. Jiang, L. J. Xie and Z. G. Yang, "Impact of BMI on left atrial strain and abnormal atrioventricular interaction in patients with type 2 diabetes mellitus: A cardiac magnetic resonance feature tracking study," *Journal of Magnetic Resonance Imaging,* 2022.

[40] L. Cheng, H. Zhuang, H. Ju, S. Yang, J. Han, R. Tan and Y. Hu, "Exposing the causal effect of body mass index on the risk of type 2 diabetes mellitus: a mendelian randomization study.," *Frontiers in genetics,* vol. 10, no. 94, 2019.

[41] B. Mi, C. Wu, X. Gao, W. Wu, J. Du, Y. Zhao and H. Yan, "Long-term BMI change trajectories in Chinese adults and its association with the hazard of type 2 diabetes: Evidence from a 20-year China Health and Nutrition Survey," *BMJ Open Diabetes Research and Care,* vol. 8, no. 1, p. e000879, 2020.

[42] A. Dagliati, S. Marini, L. Sacchi, G. Cogni, M. Teliti, V. Tibollo and R. Bellazzi, "Machine learning methods to predict diabetes complications," *Journal of diabetes science and technology,* vol. 12, no. 2, pp. 295-302, 2018.

[43] G. Maskarinec, S. Jacobs, S. Y. Park, C. A. Haiman, V. W. Setiawan, L. R. Wilkens and L. Le Marchand, "Type II diabetes, obesity, and breast cancer risk: the Multiethnic Cohort," *Cancer Epidemiology and Prevention Biomarkers,* vol. 26, no. 6, pp. 854-861, 2017.

[44] J. Luo, A. Hodge, M. Hendryx and J. E. Byles, "BMI trajectory and subsequent risk of type 2 diabetes among middle-aged women," *Nutrition, Metabolism and Cardiovascular Diseases,* vol. 31, no. 4, pp. 1063-1070, 2021.

[45] D. H. Lee, N. Keum, F. B. Hu, E. J. Orav, E. B. Rimm, W. C. Willett and E. L. Giovannucci, "Comparison of the association of predicted fat mass, body mass index, and other obesity indicators with type 2 diabetes risk: two large prospective studies in US men and women," *European journal of epidemiology,* vol. 33, no. 11, pp. 1113-1123., 2018.

[46] R. Zare, A. Nadjarzadeh, M. M. Zarshenas, M. Shams and M. Heydari, "Efficacy of cinnamon in patients with type II diabetes mellitus: A randomized controlled clinical trial," *Clinical nutrition,* vol. 38, no. 2, pp. 549-556, 2019.

[47] P. Ghosh, S. Azam, A. Karim, M. Hassan, K. Roy and M. Jonkman, "A comparative study of different machine learning tools in detecting diabetes," *Procedia Computer Science,* vol. 192, pp. 467-477, 2021.

[48] A. Qayyum, S. Talpur and M. Jawaid, "Early Detection of Type 2 Diabetes using supervised machine learning," *Engineering Science and Technological International Research Journal,* vol. 1, p. 5, 2021.

[49] I. D. Oladipo and A. O. Babatunde, "Framework for genetic-neuro-fuzzy inferential system for diagnosis of diabetes mellitus," *Annals Comput. Sci. Series,* vol. 16, no. 1, 2018.

[50] S. K. Mohapatra, J. K. Swain and M. N. Mohanty, "Detection of diabetes using multilayer perceptron," in *International conference on intelligent computing and applications*, Singapore, 2019.

[51] O. Banerjee and K. V. V. Satyanarayana, "Prediction of Diabetes Mellitus using Ensembled Machine learning Techniques," *Annals of the Romanian Society for Cell Biology,* pp. 701-711, 2021.

[52] T. A. Assegie, T. Karpagam, R. Mothukuri, R. L. Tulasi and M. F. Engidaye, "Extraction of human understandable insight from machine learning model for diabetes prediction," *Bulletin of Electrical Engineering and Informatics,* vol. 11, no. 2, pp. 1126-1133, 2022.

[53] Q. M. Yas, "Evaluation Multi Diabetes Mellitus Symptoms by Integrated Fuzzy-based MCDM Approach," *Turkish Journal of Computer and Mathematics Education (TURCOMAT),* vol. 12, no. 13, pp. 4069-4082, 2021.

[54] M. M. Bukhari, B. F. Alkhamees, S. Hussain, A. Gumaei, A. Assiri and S. S. Ullah, "An improved artificial neural network model for effective diabetes prediction," *Complexity,* 2021.

[55] Z. Mushtaq, M. F. Ramzan, S. Ali, S. Baseer, A. Samad and M. Husnain, "Voting Classification-Based Diabetes Mellitus Prediction Using Hypertuned Machine-Learning Techniques," *Mobile Information Systems,* 2022.

[56] N. Kumar, N. Narayan Das, D. Gupta, K. Gupta and J. Bindra, "Efficient automated disease diagnosis using machine learning models," *Journal of Healthcare Engineering,* 2021.

[57] S. You and M. Kang, "A Study on Methods to Prevent Pima Indians Diabetes using SVM," *Korea Journal of Artificial Intelligence,* vol. 8, no. 2, pp. 7-10, 2020.

[58] R. Barhate and P. Kulkarni, "Analysis of classifiers for prediction of type ii diabetes mellitus," *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA),* no. IEEE., pp. 1-6, 2018.

[59] S. Larabi-Marie-Sainte, L. Aburahmah, A. R. and T. Saba, "Current techniques for diabetes prediction: review and case study," *Applied Sciences,* vol. 21, no. 4604, p. 9, 2019.

[60] Y. Liu, Z. Zhao, J. Wang, A. Li and J. Zhang, "Research on Diabetes Management Strategy Based on Deep Belief Network," *International Conference on Wireless and Satellite Systems,* no. Springer, Cham., pp. 177-186, 2019.

[61] B. Farhana, K. Munidhanalakshmi and R. M. Mohana, "Predict Diabetes Mellitus Using Machine Learning Algorithms," *Journal of Physics: Conference Series ,* vol. 2089, no. 1, p. 012002, 2021.

[62] C. Roversi, E. Tavazzi, M. Vettoretti and B. Di Camillo, "A Dynamic Bayesian Network model for simulating the progression to diabetes onset in the ageing population," *2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI) ,* vol. IEEE, pp. 1-4, 2021.

[63] D. Khangura, L. R. Kurukulasuriya, A. Whaley-Connell and J. R. Sowers, "Diabetes and hypertension: clinical update," *American Journal of Hypertension,* vol. 31, no. 5, pp. 515-521, 2018.

[64] A. Massaro, A. V. Maritati, D. Giannone, D. Convertini and A. Galiano, "LSTM DSS Automatism and Dataset Optimization for Diabetes Prediction," *Applied Sciences,* vol. 9, no. 3532, p. 17, 2019.

[65] A. Massaro, G. Meuli, N. Savino and A. Galiano, "Voice Analysis Rehabilitation Platform based," *International Journal of Telemedicine and Clinical Practices (IJTMCP),* vol. 3, no. 4, 2022.

[66] A. Massaro, V. Maritati, N. Savino, A. Galiano, D. Convertini, E. De Fonte and M. Di Muro, "A Study of a Health Resources Management Platform Integrating Neural Networks and DSS Telemedicine for Homecare Assistance," *Information,* vol. 9, no. 176, pp. 1-20, 208.

[67] A. Massaro, V. Maritati, N. Savino and A. Galiano, "Neural Networks for Automated Smart Health Platforms oriented on Heart Predictive Diagnostic Big Data Systems," *IEEE Proceeding AEIT,* 2018.

[68] A. Galiano, A. Massaro, B. Boussahel, D. Barbuzzi, F. Tarulli, L. Pellicani, L. Renna, A. Guarini, G. De Tullio, G. Nardelli, R. Bonaduce, C. Minoia, S. Ciavarella, V. De Fazio, A. Negri and C. Marchionna, "Improvements in Haematology for Home Health Assistance and Monitoring by a Web based Communication System," *Proceeding of IEEE International Symposium on Medical Measurements and Applications (MeMeA),* 2016.

[69] A. Massaro, A. Galiano, D. Scarafile, A. Vacca, A. Frassanito, A. Melaccio and F. Attivissimo, "Telemedicine DSS-AI Multi Level Platform for Monoclonal Gammopathy Assistance," *IEEE Proceeding of MeMeA 2020,* 2020.

[70] A. Massaro, G. Ricci, S. Selicato, S. Raminelli and A. Galiano, "Decisional Support System with Artificial Intelligence oriented on Health Prediction using a Wearable Device and Big Data," *2020 IEEE International Workshop on Metrology for Industry 4.0 & IoT,* pp. 718-723, 2020.

[71] A. Massaro, Electronic in Advanced Research Industry: From Industry 4.0 to Industry 5.0 Advances, Wiley/IEEE, 2021.

[72] A. Massaro, N. Magaletti, V. O. Giardinelli, G. Cosoli, A. Leogrande and F. Cannone, "Original Data Vs High Performance Augmented Data for ANN Prediction of Glycemic Status in Diabetes Patients," *University Library of Munich, German,* no. 112638, 2022.

[73] A. Massaro, V. O. Giardinelli, G. Cosoli, N. Magaletti and A. Leogrande, "The Prediction of Hypertension Risk," *University Library of Munich, Germany,* no. 113242, 2022.

[74] A. Massaro, N. Magaletti, G. Cosoli, A. Leogrande and F. Cannone, "Use of Machine Learning to Predict the Glycemic Status of Patients with Diabetes," 2021.

[75] C. J. van der Kallen, G. J. Biessels and C. D. Stehouwer, "The Role of Hyperglycemia, Insulin Resistance, and Blood Pressure in Diabetes-Associated Differences in Cognitive PerformancedThe Maastricht Study," *Diabetes Care,* vol. 40, no. 1537, 2017.