sciendo

ARS Medica Tomitana - 2024; 2(30): pag. 38 - 55
doi: 10.2478/arsm-2024-0007

STATISTICAL AND MACHINE LEARNING-BASED PREDICTIVE MODELS FOR GESTATIONAL DIABETES MELLITUS PREVENTION

Hanane Zermane[1], Adel Kalla[2]

[1] Laboratory of Automation and Manufacturing, Industrial Engineering Department, Batna 2 University, Batna, Algeria
[2] Laboratory of Biotechnology of Bioactive Molecules and Cellular Pathophysiology, Microbiology and Biochemistry Department, Batna 2 University, Batna, Algeria

Hanane Zermane
email: h.zermane@univ-batna2.dz

ABSTRACT

The focus of this paper is to use machine learning to create predictive models that detect the probable factors impacting Gestational Diabetes Mellitus (GDM) which is developed in some pregnant women. GDM is defined as any proportion of glucose intolerance developed during pregnancy. Several factors may cause GDM complications. Here, we aimed to identify factors predisposing to GDM and predict the occurrence based on several predictive models. The dataset used in this study is the Pima Indian. With the assistance of Machine Learning and Statistical Analysis, it is possible to develop intelligent models that are capable of making decisions on an autonomous basis. Seven machine learning models were tested to determine which model fits the dataset better. These models learn from past instances of data through Statistical Analysis and pattern matching. Based on the learned data, they provide us with the predicted results. This study establishes the feasibility of machine learning in the field of public health. It is observed that each technique gives different results of associated factors. The Cascade classifier model attained an accuracy of 98.58%, Random Forest (89%), SVM (69%), Logistic Regression (78%), K-NN (72%), and Decision Tree (78%). These models are validated and evaluated using several metrics. This work demonstrated that identifying risk factors must not consider one model.

Keywords: Gestational Diabetes Mellitus, Data Analysis, Predictive Models, Ascending Hierarchical Clustering, Multiple Corresponding Analysis, Machine Learning, Cascade Forest Classifier, Random Forest Classifier.

Introduction

World Health Organization (WHO) defines diabetes as a disorder that occurs when the pancreas does not create enough insulin or when the body does not utilize the insulin that is produced adequately. The most common types of diabetes are type 1, type 2, and Gestational Diabetes Mellitus (GDM) (1). GDM is developed in some pregnant women. It is defined as any proportion of glucose intolerance developed during pregnancy and is related to higher maternal and fetal mortality as well as long-term problems in both mothers and babies (2). It is known as hyperglycemia with onset or first confirmation during pregnancy. Subclinical hyperglycemia was associated with impaired insulin sensitivity in women with a history of GDM (3). It has symptoms that are analogous to Type 2 diabetes (4). Prenatal screening, rather than reported symptoms, is the most common way to identify GDM (5). Most of the time, this type of diabetes goes away after the baby is born. However, if GDM is taking place, it will develop into type 2 diabetes later in life. Occasionally, diabetes diagnosed during pregnancy is type 2 diabetes (6). To prevent this critical pathology and save a mother's life, the prediction of its happening is crucial. One of the most efficient and satisfactory techniques, including statistical

analysis and artificial intelligence (AI) with belonging techniques such as Machine Learning (ML) techniques. Building predictive models to achieve this objective is an improved and unavoidable solution. By bridging gaps between massive datasets and human understanding, Machine Learning is assisting medical practitioners in providing diagnoses simpler, faster, and better by predicting associated factors of a given risk. Machine Learning is utilized to extract correlations and dependencies of the data and use them either to understand a phenomenon or to predict future outcomes. However, it can be a very complex problem, and there is usually no perfect method that can give the correct label every time. However, in most cases, a good prediction, even if it is imperfect, is still better than no prediction at all.

Accordingly, data analysts and experts must not consider just one technique or create one model. Diversity can make the change and diversify choices to make a good decision. Sometimes it is possible to produce better predictions on our own, but it is still preferred to use Statistics and Machine Learning because the machine will make its predictions faster and with improved precision. Powerful models can be built as software helping the medical team by reducing diagnosis time and improving their decision about any studied population, even if the amount of data is massive.

Intelligent models have proven their importance, in trying to solve the various problems that can arise. This can help to make a decision by specialists in the medical domain and try to achieve intelligent behavior with computers, through the various techniques of Statistics and Machine Learning. ML enables models to be trained on datasets before deployment. Machine Learning models increase the types of correlations created between data pieces through an iterative process. A human may not be able to distinguish these patterns and relationships due to their complexity and magnitude. Once a model is trained, it may be utilized to learn from data in real-time. The training processes result in increased accuracy.

In this context and based on the improved advantages of Statistical Analysis techniques, including Multivariate Analysis, The Generalized Linear Model, Ascending Hierarchical Clustering, and Multiple Corresponding Analyses. Concerning Machine Learning Techniques, the Deep Cascade Forests and Random Forests techniques are applied to solving classification problems to predict the occurrence of gestational diabetes mellitus. The two models are validated and evaluated by comparison to other ML models, namely Support Vector Machine (SVM), Logistic Regression, Decision Trees, and K-Nearest Neighbor (K-NN).

Materials

Data preparation involves transforming the raw data into a form more suitable for modeling. It may be the most important part of a predictive modeling project and the longest. Instead, the emphasis is on machine learning algorithms, which have become quite routine to use and configure. Practical data preparation requires knowledge of data cleaning, feature selection, data transformations, dimensionality reduction, and more. Features selection is the main core of the prediction, diagnosis, classification, clustering, and detection. These methods are used to get as much information as possible and this is the essential step in the learning process. They solve the problem of finding the most compact and informative set of features to improve efficiency in data processing. In the utilized dataset (Pima Indian dataset), about 768 patients' data are collected between the 1960s and 1980s, with 268 positives for diabetes (1) and 500 negatives for diabetes (0). About eight factors are selected to predict the target which is the outcome of having gestational diabetes or not. The Gila River Indian population is a community in Southern Arizona within a community married for 2000 years. Many Pima Indian women are diabetic. Patients are volunteers for the genetic research study made by the United States National Institutes of Health (NIH). This dataset belongs to the National Institute of Diabetes, Digestive, and Kidney Diseases.

The descriptive analysis consisted of several factors including the number of pregnancies, Age, Glucose, Blood pressure, Skin Thickness, Insulin, BMI, and Diabetes Pedigree Function. The Diabetes Pedigree Function

presents a function-scored likelihood of diabetes based on family history, it provides some data on diabetes mellitus history in relatives and the genetic relationship of those relatives to the patient, with a measurement of the degree of implication of diabetes in family history, the patient response to a know Universal questionary by endocrinologists, the final score of responses determines the Diabetes Pedigree Function. This measure of genetic influence gave us information about the hereditary risk one might have with the onset of diabetes mellitus (7). The variance between different predictor factors varied on a large scale. The scaling data will be beneficial for predictive modeling. The purpose of the dataset is to diagnose predictably whether a patient has diabetes or not, based on certain diagnostic measures included in the dataset. Several constraints were placed on the selection of these instances from a larger dataset. In particular, all of the patients here are women of at least 21 years of Pima Indian descent. Therefore, several predictive models are built to accurately predict whether the patients will possess gestational diabetes or not. The factors are collected in Table 1.

Table 1. Risk factors associated with gestational diabetes mellitus.

Risk Factors	Description
Pregnancies	Number of pregnancies
Glucose	Plasma glucose concentration at two hours in an oral glucose tolerance test
Blood Pressure	Diastolic blood pressure (hypertension) (mm Hg)
Skin Thickness	The thickness of the skin fold of the triceps (mm)
Insulin	Two-hour serum insulin (mu U/ml)
Age	Age (years)
BMI	Body Mass Index (kg/m2)
Diabetes Pedigree Function	A measure of genetic influence gave us information about the hereditary risk one might have with the onset of diabetes mellitus (Pedi)

Methods

To discriminate between diabetes-affected and not-affected women by developing a predictive model, the Pima Indian dataset provided a well-validated data resource in which to explore the prediction of the date of onset of diabetes in a longitudinal manner. That population of volunteer patients was under continuous study from 1960 to 1980 because of its high prevalence rate of diabetes. Each community resident over five years of age was requested to experience a standardized inspection every two years. The diagnosis was made by the medical staff through convolutional methods of diabetes diagnosis including an oral glucose tolerance test (8). Thus, the a need to utilize complementary approaches with sophisticated techniques that allow the prevention of such diseases and identifying associated factors such as Statistical Analysis and Artificial Intelligence techniques. Selecting the best classifier for the objectives of a specific situation is, therefore, a real problem. There are many considerations to take into account, ranging from the type of data available to the objectives of the decision-maker, including the cost of potential errors. In this work, Several Statistical Analysis techniques are performed, including Univariate and Multivariate Analysis, The Generalized Linear Model, Ascending Hierarchical Clustering, and Multiple Correspondence Analysis. Moreover, there are many classifiers, each with its strengths and weaknesses. Among the best known, Support Vector Machine (SVM), Logistic Regressions, Classification and Regression Trees (CART Trees), Random Forests, and Deep Cascade Forests are cited. We are in favor of using the Random Forest and Cascade Forests techniques in this work. These techniques achieved higher prediction performance than conventional techniques. Due to their benefits in comparison to other statistical methods, Random Forest is a popular tool in a wide range of sectors, especially the medical field, however, the Deep Cascade Forest is newly been integrated.

Statistical Analysis Techniques

Statistical and descriptive techniques have proven their importance and justified their use by researchers in various fields (9–11). Logistic regression techniques are prized in epidemiology to identify the factors associated with this or that pathology. Univariate and multivariate logistic regression models were used to assess the effect of covariates on the odds ratios (OR) of factors among risks. Potential confounders with a p-value < 0.20 in univariate analysis were retained for the

final multivariate analysis. Logistic Regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analysis techniques, logistic regression is a predictive analysis used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval, or ratio-level independent variables (12). Logistic Regression (LR) calculates the class membership probability for two categories by fitting the log odds and explanatory variables to the identified model (13). Logistic Regression provides a model of logistic probability. It is easy to interpret, it provides confidence intervals and it quickly updates the classification model to incorporate new data (14). The Logistic Regression technique is applied in several studies including in the medical field (11,15).

The top-down Stepwise strategy is used to arrive at a final model that contains only predictive variables whose association with the response variable is significant or close to significant. By applying Stepwise, it is always tempting when looking for associated factors to include a large number of potential explanatory variables in the model (16). However, such a model is not necessarily the most efficient, and some factors will probably not have a significant effect on the risk. The need to use other techniques is crucial. The most utilized are Multiple Correspondence Analysis and Ascending Hierarchical Clustering. The ascending Hierarchical Clustering technique is one of the statistical techniques aiming at dividing a population into different classes or subgroups. It is utilized to seek that the individuals grouped within the same class (inter-class homogeneity) are as similar as possible while the classes are the most dissimilar (inter-class heterogeneity). Consequently, it allows us to determine the presence of homogenous groups of more highly correlated subjects or variables. Multiple Correspondence Analysis (MCA) is a descriptive technique that is related to several qualitative variables. The MCA technique summarizes the information contained in a large number of variables to facilitate the interpretation of the correlations existing between these different variables and try to find out which modalities are correlated with each other (12).

Machine Learning Techniques

The classification task is based on supervised learning. Its principle is to build a classification model capable not only of describing the class of individuals classified a priori but also of predicting the class of new individuals not classified a priori. The classes, in this case, are discrete values (absence or presence of a disease). Usually, the technique used in classification is called a classifier. It is the primary tool utilized to perform predictive analytics tasks. ML has the benefit of allowing forecasting outcomes using algorithms and models (17). The key is to ensure that data scientists are utilizing the most appropriate algorithms and executing models after data processing and cleaning. If all of these aspects work together, the model may be constantly trained and the outcomes exploited by learning from the data. Support Vector Machine (SVM) developed during the 90s by Vapnik (18,19) are a family of learning algorithms intended to solve classification and regression problems and the detection of new values. SVM is today considered one of the most efficient methods for many real problems, in particular for large dimension problems. The effectiveness of SVM is due to its solid theoretical basis which is one of its strengths (20). SVMs have several advantages (21), it is efficient in large spaces and always effective in cases where the number of dimensions exceeds the number of samples. It uses a subset of learning points in the decision function (support vectors), so it is also memory efficient. It is versatile with different kernel functions that can be specified for the decision function. Common kernels are provided, but it is also possible to specify custom kernels. SVM is applied in several studies in different other fields (22,23), especially, in medicine and health (24–26).

Nearest Neighbor (K-NN) is a non-parametric discrimination method, no parameter estimation is necessary for its execution. It works on continuous data. The main idea is to observe the K closest to a new observation to decide on the membership of this observation. To classify an observation, the distance between the new observation and each point in the learning base is calculated. Afterward, the K neighbors having the smallest distance from the new observation

will be selected. Because of the membership classes of the K-nearest neighbors, it is decided on the membership class of the new observation (27). The Nearest Neighbor technique has been used for several purposes (22,27).

Decision trees are a category of trees used in data mining and business intelligence. They employ a hierarchical representation of the data structure in the form of sequences of decisions (tests) for the prediction of an outcome or a class. Each instance, which must be assigned to a class, is described by a set of factors that are tested in the nodes of the tree. Testing takes place in internal nodes and decisions are made in leaf nodes (24,28). Trees have many advantages but tend to overfit the data, which affects their accuracy when utilized in production mode. Random Forests classifiers are made up of a set of trees of different sizes and shapes, which remedy this problem. It protected against overfitting and detected interactions between variables. In addition, these algorithms have significantly improved predictive capabilities (29,30), giving an error rate that is generally lower, and at worst comparable, to that of most other classification methods invented to date. The first forest-type classifier was the Random Forest introduced by L. Breiman (31). It is a set version of CART trees and is one of the most accurate and efficient methods on the market. L. Breiman defined a Random Forest as a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest (31). In the Random Forest model, the number of trees that vote for an outcome gave a statistical probability that an input factor is related to an output factor. It is, therefore, a voting method based on the assumption that most methods are correct most of the time and predicts the output class that will receive the maximum number of votes from all individual decision trees (26). Unlike decision trees, random forests are less likely to over-adapt to training data.

The Deep Cascade forests (known as gcForest) technique is a deep Forests algorithm suggested by Zhou and Feng in 2017 (32). It uses a multi-grain scanning approach for data slicing and a cascade structure of multiple random forest layers. The first gcForest implementation has been developed as a classifier and designed such that the multi-grain scanning module and the cascade structure can be used separately. A deep forest is a Deep Neural Network (DNN), it may open a door toward an alternative to deep neural networks for many tasks. Unlike deep neural networks, it is composed of differentiable neurons. The basic components of Deep Cascade Forests are nondifferentiable decision trees. The process of training does not depend on gradient calculation. It preliminarily verifies the conjecture about the reason why deep learning works. The complexity of the model is enough to build an effective deep-learning model. Deep Forests' main characteristics include better performance than other ensemble learning methods based on decision trees with fewer hyperparameters, it does not need a lot of tuning, high training efficiency, and the possibility of handling large datasets.

Results

In these experiments, as a programming language, Python v3.9.7 and R v4.2.1 (under RStudio Desktop 2022.07.1+554) are performed to analyze data and develop the predictive models using an Asus laptop, with a Core i7 processor, graphical card of 2 GByte, and 16 GByte of RAM. The data is biased towards data points with an outcome value of "0", which means that gestational diabetes was not present. The number of diabetic patients (500) is almost double the number of healthy which is 268. The descriptive analysis consisted of several factors including the number of pregnancies, Age, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, and Diabetes Pedigree Function. The prevalence of diabetic and healthy women in this population is gathered in Table 2. Reported statistics prove the high prevalence for Age > 35 (>47%), BMI > 30 (45.2%), Blood Pressure > 80 (46.7%), Pregnancies > 6 (>56%), Insulin > 200 (>52%), Diabetes Pedigree Function > 0.825 (>52%), Skin Thickness > 54 and Glucose >150 (75%).

Table 2. Diabetes prevalence according to different risk factors.

Risk Factors	Healthy		Diabetic	
	N°	Prevalence (%)	N°	Prevalence (%)
Age [21,35]	367	73.7	131	26.3
Age (35,50]	90	47.6	99	52.4
Age (50,81]	43	53.1	38	46.9
BMI (19,25]	101	93.5	7	6.5
BMI (25,30]	136	75.6	44	24.4
BMI (30,35]	121	54.8	100	45.2
Blood Pressure (0,72]	275	71.6	109	28.4
Blood Pressure (72,80]	118	64.1	66	35.9
Blood Pressure (80,122]	88	53.3	77	46.7
Pregnancies (0,2]	190	79.8	48	20.2
Pregnancies (2,6]	163	65.2	87	34.8
Pregnancies (6,10]	60	44.4	75	55.6
Pregnancies (10,17]	14	41.2	20	58.8
Insulin (80,200]	129	63.2	75	36.8
Insulin (200,300]	23	47.9	25	52.1
Insulin (300,400]	8	47.1	9	52.9
Insulin (400,850]	8	40.0	12	60.0
Diabetes Pedigree Function (0.078,0.33]	246	73.0	91	27.0
Diabetes Pedigree Function (0.33,0.825]	207	61.8	128	38.2
Diabetes Pedigree Function (0.825,2.42]	46	48.4	49	51.6
Skin Thickness (0,11]	14	93.3	1	6.7
Skin Thickness (11,54]	346	66.3	176	33.7
Skin Thickness (54,99]	1	25.0	3	75.0
Glucose (0,95]	143	92.3	12	7.7
Glucose (95,110]	130	81.2	30	18.8
Glucose (110,150]	189	61.4	119	38.6
Glucose (150,199]	35	25.0	105	75.0

To highlight the correlation and assess the dependence of different factors at the same time, a heatmap matrix is utilized. The matrix contains the correlation coefficients between each factor and the others. This correlation is very often reduced to the linear correlation between quantitative factors, the adjustment of one factor with the other by an affine relation obtained by linear regression. Accordingly, for a linear correlation coefficient, the quotient of their covariance by the product of their standard deviations is calculated. Its sign indicates whether higher values of one correspond on average to higher or lower values for the other.

The absolute value of the coefficient is in the range between [0 - 1]. This value does not measure the strength of the link but the preponderance of the affine relation over the internal variations of the factors. A zero coefficient does not imply independence. Other indicators allow the calculation of a correlation coefficient for ordinal factors. The fact that the two factors are strongly correlated does not demonstrate that there is a causal relationship between them. The most typical counterexample is where they are linked by a common causality. Accordingly, information that could be extracted from the heatmap (see Figure 1) indicates that the most correlated factors to diabetes are Glucose, BMI, Age, number of Pregnancies, and after that, Diabetes Pedigree Function, and Insulin.

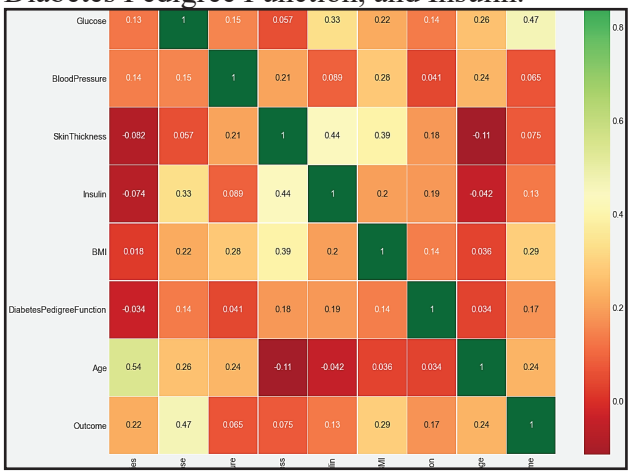

Figure 1 Correlation between diabetes and associated risk factors.

The distribution of the number of diabetic and non-diabetic women according to each risk factor is displayed in Figure 2. It is observed that the number of pregnancies in the population of patients increased the probability of having gestational diabetes.

The importance of factors refers to a group of methods for awarding scores to inputs. The relative relevance of each component in a forecast is indicated by factors in a predictive model. For issues requiring forecasting a numeric value, the regression is carried out, and for problems involving predicting a class label, the classification is performed, and factor significance scores can be determined. Scores are valuable and may be employed in a predictive modeling challenge for a variety of reasons, including greater data interpretation, a better

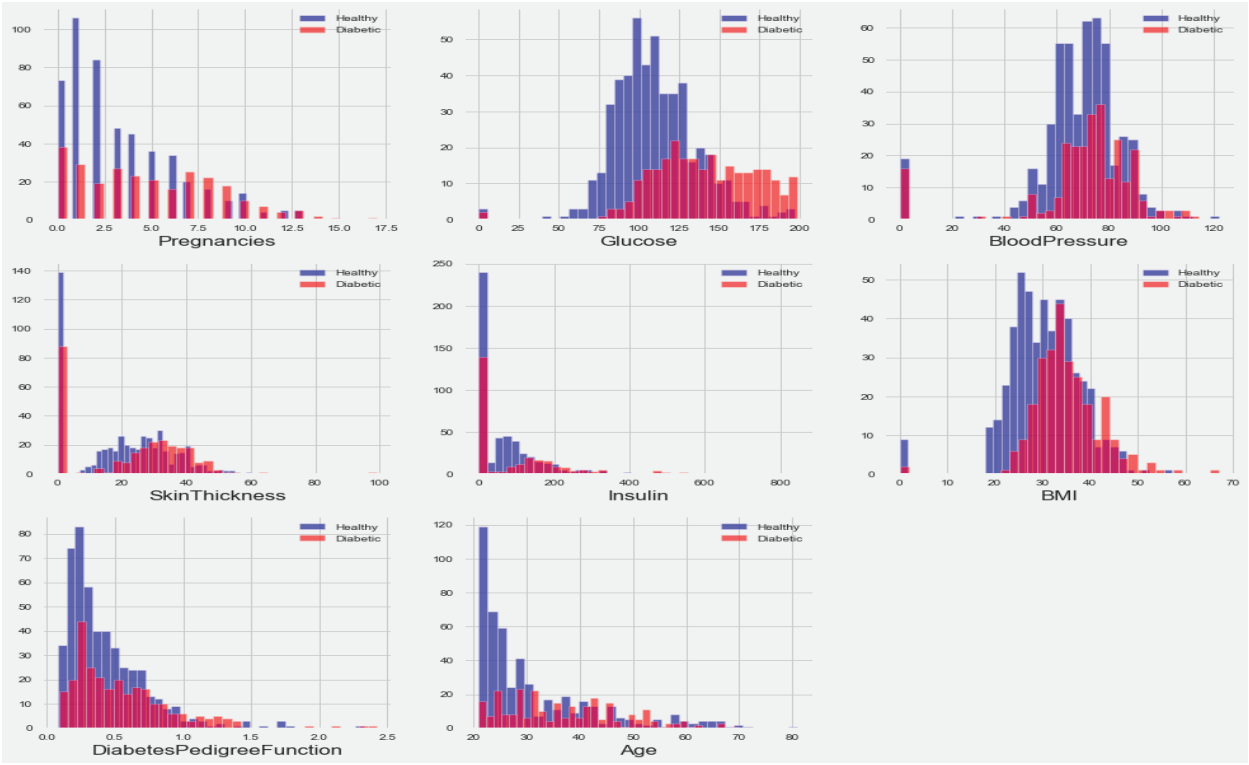

Figure 2. Distribution of diabetic and healthy patients according to each risk factor.

understanding of the model, and a decrease in the number of input entities. Based on the graph of age distribution, there may not be a strong correlation between age and the onset of gestational diabetes. However, it is noticeable that an association exists between Glucose and diabetes.

Statistical analysis of diabetes and risk factors associations

To highlight primary correlations and potential associations of risk factors previously illustrated in Table 2, a univariate analysis of the odds ratio (interval of confidence IC:97%) is utilized. However, the association may be biased by the univariate aspect of the analysis and the influence of the combination of risk factors could not be known. Therefore, multivariate analysis is carried out. The univariate and bivariate analysis results are reported in Table 3. It is observed that influential factors include, BMI > 25 (p=0.055, p=0.015), Blood Pressure > 80 (p=0.048), Insulin > 400 (p=0.136), Diabetes Pedigree Function > 0.33 (p=0.187, p=0.006), and Glucose > 150 (p=0.013). However, based on this multivariate analysis with adjusted odds ratios, factors such as Age, Skin Thickness, and number of Pregnancies do not influence diabetes.

Based on previous analysis, it is hard to validate an obvious association between factors (Age, BMI, Blood Pressure, Glucose, Insulin, Diabetes Pedigree Function, Pregnancies, and Skin Thickness) and their influence. Therefore, to adjust a potential statistical co-founding bias, a multivariate logistic regression analysis is performed. Possible interactions between factors were established including all factors in a complete model utilizing a Generalized Linear Model (GLM) (see Table 4). It is observed that factors including BMI > 25, Blood Pressure > 80 (*p=0.05402, p=0.01726*), Glucose > 150 (*p=0.0187*), Insulin > 400 (*p=0.16522*), and Diabetes Pedigree Function > 0.33 (*p=0.18300, p=0.00408*) are the most influential factors. While, factors less important include, Age, Blood Pressure [72,80], Glucose < 150, Insulin < 400, Number of Pregnancies, and Skin Thickness with an AIC of 176.82.

Table 3. Odds Ratio (OR, 97% CI) of risk factors potentially associated with Diabetes.

Risk factors range	OR (univariable)			OR (multivariable)			adjusted OR (final model)		
	OR	97% IC	p-value	AOR	97% IC	p-value	AOR	97% IC	p-value
Age (35,50]	3.08	(2.18-4.37)	p<0.001	1.20	(0.34-4.08)	p=0.767			
Age (50,81]	2.48	(1.53-4.00)	p<0.001	1.63	(0.24-11.80)	p=0.618			
BMI (25,30]	4.67	(2.14-11.73)	p<0.001	10.70	(1.42-261.36)	p=0.054	10.18	(1.40-241.96)	p=0.055
BMI (30,35]	11.92	(5.66-29.32)	p<0.001	19.71	(2.58-500.97)	p=0.017	19.35	(2.68-465.00)	p=0.015
Blood Pressure (72,80]	1.41	(0.97-2.05)	p=0.071	1.39	(0.43-4.37)	p=0.577	1.42	(0.44-4.42)	p=0.548
Blood Pressure (80,122]	2.21	(1.51-3.22)	p<0.001	3.63	(0.93-15.15)	p=0.067	3.77	(1.03-14.74)	p=0.048
Pregnancies (2,6]	2.11	(1.41-3.20)	p<0.001	1.12	(0.42-2.99)	p=0.815	1.11	(0.42-2.90)	p=0.833
Pregnancies (6,10]	4.95	(3.13-7.92)	p<0.001	1.66	(0.33-8.35)	p=0.535	2.07	(0.56-7.72)	p=0.272
Pregnancies (10,17]	5.65	(2.69-12.23)	p<0.001	4.72	(0.32-87.23)	p=0.262	4.36	(0.37-63.41)	p=0.245
Insulin (200,300]	1.87	(0.99-3.54)	p=0.053	0.55	(0.13-2.19)	p=0.408	0.51	(0.12-1.95)	p=0.339
Insulin (300,400]	1.93	(0.71-5.36)	p=0.193	0.55	(0.09-3.06)	p=0.508	0.51	(0.08-2.73)	p=0.443
Insulin (400,850]	2.58	(1.02-6.86)	p=0.048	11.68	(0.65-762.36)	p=0.165	12.57	(0.83-731.30)	p=0.136
DPF(0.33,0.825]	1.67	(1.21-2.32)	p=0.002	1.93	(0.75-5.26)	p=0.183	1.90	(0.75-5.08)	p=0.187
DPF (0.825,2.42]	2.88	(1.80-4.61)	p<0.001	7.93	(2.03-35.14)	p=0.004	7.02	(1.83-30.17)	p=0.006
Skin Thickness (11,54]	7.12	(1.41-129.58)	p=0.059	3.45	(0.34-95.86)	p=0.355			
Glucose (95,110]	2.75	(1.38-5.79)	p=0.005	0.41	(0.07-2.68)	p=0.325	0.42	(0.07-2.73)	p=0.338
Glucose (110,150]	7.50	(4.14-14.81)	p<0.001	1.28	(0.28-7.36)	p=0.761	1.36	(0.30-7.73)	p=0.705
Glucose (150,199]	35.75	(18.33-75.30)	p<0.001	8.02	(1.54-52.88)	p=0.019	9.00	(1.75-59.21)	p=0.013

***Notes**. OR: Odds Ratio; 97%IC: Interval of Confidence at 97%; *: OR significant at α=0.1; **: OR significant at α=0.05; ***: OR significant at α=0.01.*

Table 4. The GLM of Diabetes includes all risk factors.

Risk factors	Estimate	Std. Error	z value	Pr(>|z|)
Age (35,50]	0.1863	0.6277	0.297	0.76665
Age (50,81]	0.4862	0.9745	0.499	0.61787
BMI (25,30]	2.3704	1.2303	1.927	0.05402 .
BMI (30,35]	2.9810	1.2520	2.381	0.01726 *
Blood Pressure (72,80]	0.3268	0.5860	0.558	0.57703
Blood Pressure (80,122]	1.2889	0.7049	1.828	0.06748 .
Glucose (95,110]	-0.9021	0.9169	-0.984	0.32520
Glucose (110,150]	0.2470	0.8136	0.304	0.76143
Glucose (150,199]	2.0816	0.8852	2.351	0.01870 *
Insulin (200,300]	-0.5948	0.7181	-0.828	0.40752
Insulin (300,400]	-0.5909	0.8927	-0.662	0.50798
Insulin (400,850]	2.4577	1.7710	1.388	0.16522
Diabetes Pedigree Function (0.33,0.825]	0.6571	0.4935	1.332	0.18300
Diabetes Pedigree Function (0.825,2.42]	2.0709	0.7211	2.872	0.00408 **
Pregnancies (2,6]	0.1160	0.4966	0.234	0.81523
Pregnancies (6,10]	0.5091	0.8196	0.621	0.53455
Pregnancies (10,17]	1.5524	1.3826	1.123	0.26151
Skin Thickness (11,54]	1.2397	1.3399	0.925	0.35483

***Notes**. Signification: codes : *: p<0.1, **: p<0.05, ***: p<0.001. (Dispersion parameter for binomial family taken to be 1); Null deviance: 205.89 on 155 degrees of freedom, Residual deviance: 138.82 on 137 degrees of freedom; AIC: 176.82; Number of Fisher Scoring iterations: 6.*

The significance ratings of factors can be used to get insight into the dataset. Relative scores can reveal which elements are most important to Diabetes, as well as which are least important. This can be analyzed by a medical professional and used as a starting point for gathering more or alternative data. It is always tempting when looking for the factors associated with a phenomenon to include a large number of potential explanatory variables in a logistic model. However, such a model is not necessarily the most efficient, and certain variables will probably not have a significant effect on the variable of interest. The step-by-step selection technique is an approach aimed at improving its explanatory model. It is also necessary to define a criterion to determine the quality of a model.

One of the most widely used methods of model selection is the Akaike Information Criterion (AIC). This technique allows selecting the best model of influencing factors using a Top-down Stepwise regression strategy based on minimizing AIC and returns the final model (The lower the AIC is, the better the model is). The Top-down Stepwise Strategy of Diabetes with associated risk factors shows that the most influential risk factors are BMI, Blood Pressure, Insulin, Diabetes Pedigree Function, and Glucose; with an AIC = 168.61 (see Table 5).

Table 5. Influential risk factors associated with gestational diabetes mellitus.

Coefficients:	Estimate	Std. Error	z value	Pr(>|z|)
BMI (25,30]	2.3981	1.2118	1.979	0.0478 *
BMI (30,35]	2.9096	1.1972	2.430	0.0150 *
Blood Pressure (72,80]	0.4110	0.5699	0.721	0.4708
Blood Pressure (80,122]	1.6570	0.6305	2.628	0.0085 **
Glucose (95,110]	-0.8776	0.8844	-0.992	0.3210
Glucose (110,150]	0.2347	0.7751	0.303	0.7621
Glucose (150,199]	2.2238	0.8456	2.630	0.0085 **
Insulin (200,300]	-0.5900	0.6867	-0.859	0.3902
Insulin (300,400]	-0.6681	0.8787	-0.760	0.4470
Insulin (400,850]	2.6625	1.5901	1.674	0.0940 .
Diabetes Pedigree Function (0.33,0.825]	0.6156	0.4820	1.277	0.2014
Diabetes Pedigree Function (0.825,2.42]	1.9005	0.6942	2.738	0.0061 **

***Notes**. Signification: codes : *: p<0.1, **: p<0.05, ***: p<0.001. (Dispersion parameter for binomial family taken to be 1); Null deviance: 205.89 on 155 degrees of freedom, Residual deviance: 142.61 on 143 degrees of freedom; AIC: 168.61; Number of Fisher Scoring iterations: 6.*

The odds ratio and p-values are changed giving a clear view of the strength of the risk factors' effect (Interval of Confidence at 95%). To get a better overview of the retained model, the new odds ratio adjusted are presented in a forest model (see Figure 3). Factors including; BMI > 25, Blood Pressure > 72, Pregnancies > 6, Insulin [400-850], Diabetes Pedigree Function > 0.33, and Glucose > 110, appeared on the right side of the vertical line of the forest model that represents the risk factors estimated in the population confirming their positive association. However, Insulin < 300 and Glucose < 110, appeared on the left side of the forest model, which means that these factors are less influential or negatively associated.

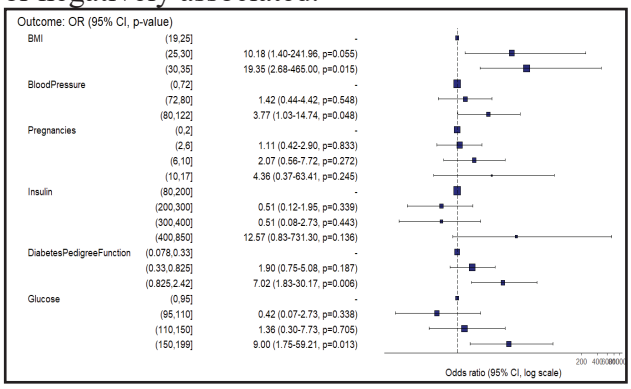

Figure 3. Forest model of strength-associated risk factors and their influence on diabetes in the predictive model.

To identify similarities among our population, an Ascending Hierarchical Clustering (AHC) is carried out. This approach contains two steps, in the first step, AHC iteratively gathered data to generate a clustering tree.

In the second step, pairs of clusters most similar are successively and iteratively merged in a dendrogram (see Figure 4). In this population, it is observed that the associated factors include Glucose [150-199], Age [35-50], Skin Thickness [40-99], Pregnancies [6-10], and finally, Blood Pressure [80-122].

A Multiple Correspondence Analysis (MCA) is performed to summarize the information contained in a large number of variables to

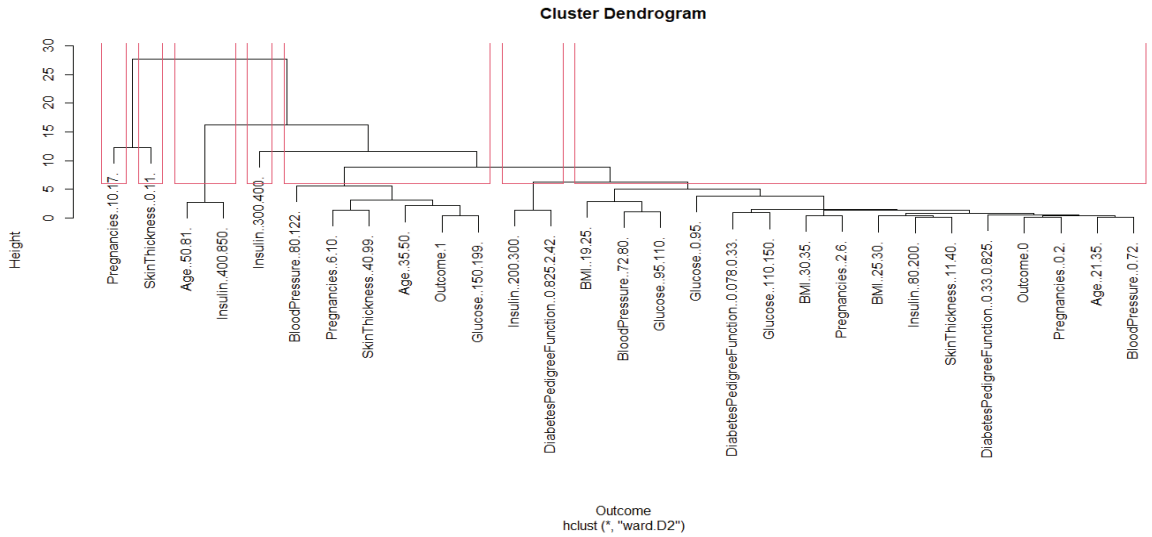

Figure 4. Cluster dendrogram of related risk factors to Diabetes.

facilitate the interpretation of the correlations existing between these different variables and find out which modalities are correlated with each other. Correspondence Analysis is all about the relativities. The proximity between levels means that the groups of observations associated with these levels are themselves similar. Points that are farther away from the origin indicate more influential categories. Points on opposite sides of the plot indicate that a component contrasts these categories. The further labeled individuals are from the origin, the more discriminating they are.

Labeled individuals that are in the center indicate that they are indistinct, and that is all that they have in common. The labeled individuals (left of Figure 5) are those with a higher contribution to the construction of the plane, while labeled variables are those best shown on the plane (right of Figure 5). In the displayed diagram (Figure 5), Dimension 1 (12.95%) and Dimension 2 (7.84%) comprised 20.79% of the total information. It is observed a small angle connecting a row and column label to the origin, they are probably associated. However, Glucose [150-199], Age [35-50], skin thickness [40-99], and pregnancies [6-10], are on the right side of the map which is Outcome 1 (diabetic). This corresponds with the relatively high contribution for these categories. Because Glucose < 110 and pregnancies < 6, are on opposite sides of the map which is Outcome 0 (healthy).

Predictive Models construction and validation

The two techniques, Random Forest and Deep Cascade Forests are detailed to illustrate the models' constructions and validations. Firstly, the Random Forest model is explained by defining all hyperparameters and validation metrics. The Random Forest classifier is a meta-estimator that fits several decision tree classifiers on various subsamples of the dataset and uses the mean to improve the predictive accuracy and control overfitting. The result of applying the Random Forests technique to the gestational diabetes dataset provided us with the evaluation metrics that demonstrate the effectiveness of the learning model. A k-fold cross-validation strategy (k=5 in our experiments) was used to test prediction models in the training set. The main advantage of this evaluation approach is that it expands the amount of data that can be used to train models

Figure 5. Distribution maps of risk factors using Multiple Correspondence Analysis.

by allowing all data instances to be used for both training and validation in different iterations. It also gives credible evaluations of prediction model performance for data that has not yet been observed. The results of the Random Forest model evaluation and validation are collected in Table 6.

Table 6. Evaluation metrics of Random Forest model.

Metric	Value
Cross-validation (5 folds)	0.940
Accuracy	0.890625
Precision	0.859375
Recall	0.820896
F1_score	0.839695

In the forest displayed in Figure 6, except for the leaf nodes (colored terminal nodes), all nodes contain five parts:

- A question regarding the data was asked depending on the value of a characteristic. Each question has a True or False answer, which divides the node. A data point travels down the tree based on the response to the query.
- Nbr of estimators: is the number of trees to use to reach the prediction (in our case n_ estimators =100).
- gini: The node's Gini impurity. As we travel down the tree, the average weighted Gini impurity drops.
- samples: The instance is the number of observations in the node that is denoted by the term samples.
- value: The number of samples in each class is denoted by the value. The top node includes two samples in class "0" and four samples in class "1".
- Class (0: healthy, 1: diabetic): The majority categorization for node points. This is the forecast for all samples in the node in the case of leaf nodes.

In ensemble classification, a majority vote from all models in the ensemble is taken, with the class that received the most votes being the final projected class. The margin thus reflects the proportion of votes cast for the right class in all instances. For one instance, a value of "1" implies that all of the ensemble's votes were right, whereas a value of "0" shows a tie between the correct and next best classes.

As a result, values > 0 indicate that the majority was accurate, and so this instance is predicted correctly, whereas all "0" values indicate that the majority is incorrect, and thus the instance is categorized incorrectly. The tree interpreter gives the sorted list of biases (mean of data at the starting node) and individual node contributions for a given prediction. The leaf nodes do not have a question because these are where the final predictions are made.

To classify a new point, simply move down

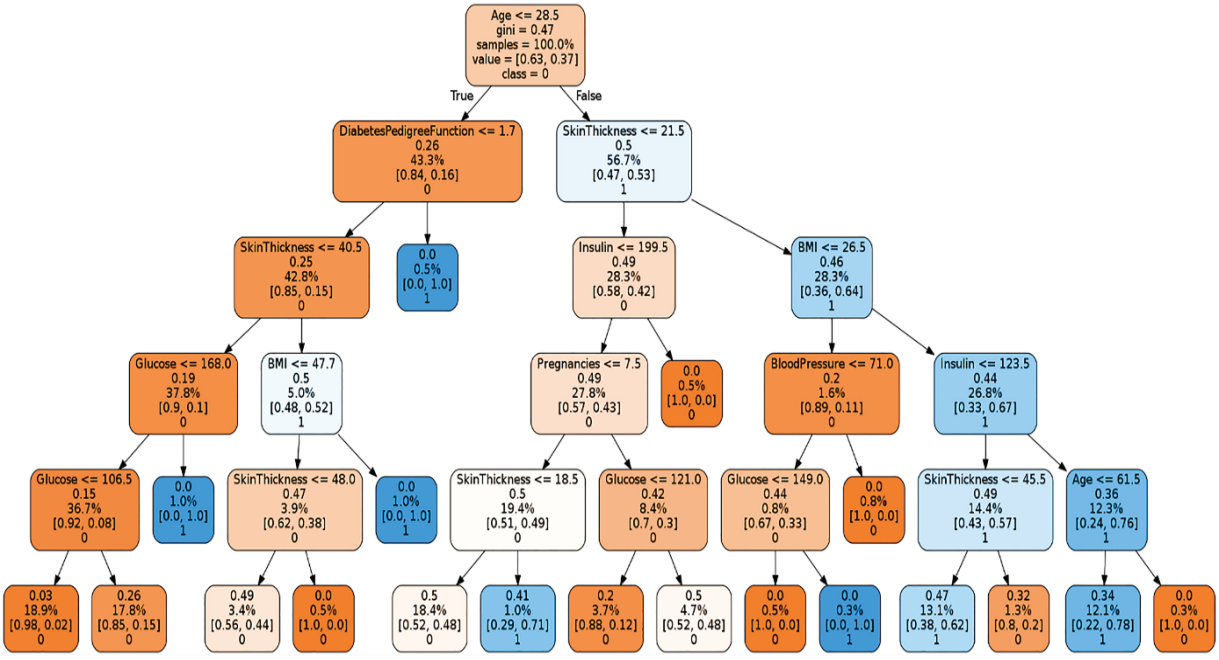

Figure 6. Example of the Random Forest of classification decision.

the tree, using the features of the point to answer the questions until you arrive at a leaf node where the class is the prediction. As an example, in the case presented in the forest displayed in Figure 6, the decision about age <= 28.5 years indicated that the patient does not have diabetes (class 0) according to the vote about the number of class "0" which is 10, and class "1" which is 4.

An additional example, in the case presented in the forest displayed in Figure 7, the decision is about 127 ≤ Glucose < 127. The predicted decision indicates that the patient does not have diabetes (class 0) according to the vote about the number of class "0" which is 5, and class "1" which is 2.

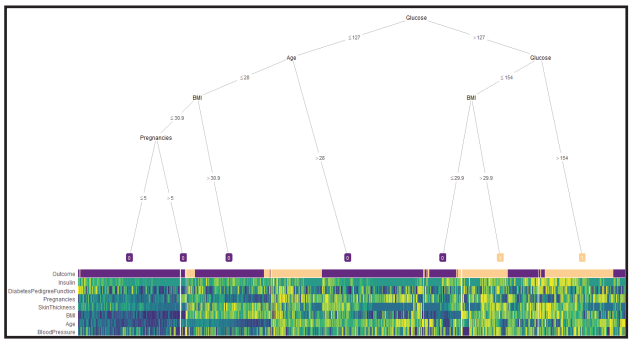

Figure 7. Random Forest of influential risk factors.

To validate the model, the Receiver Operator Characteristics (ROC) index shows the classification's overall factor performance as a discriminating threshold (relationship between specificity and sensitivity). The AUC is the classifier's general quality index. The nearest point to the upper corner, where the true positive rate is "1" and the false positive rate is "0" is point 0.8. The area under the curve (AUC) of the ROC, is the overall measure of the Random Forest classifier performance, it is calculated to compare the overall performance of all the different classification schemes.

Concerning the model validation, Random Forest is based on bootstrap methods, which are a resampling strategy commonly used to validate the model and estimate statistics on a population. Each bootstrap sample develops a predictive model, and each bootstrap sample estimates metrics of predictive performance. After that, the bootstrap models are applied to the original dataset to validate the model and establish the predictive measure (see Figure 8).

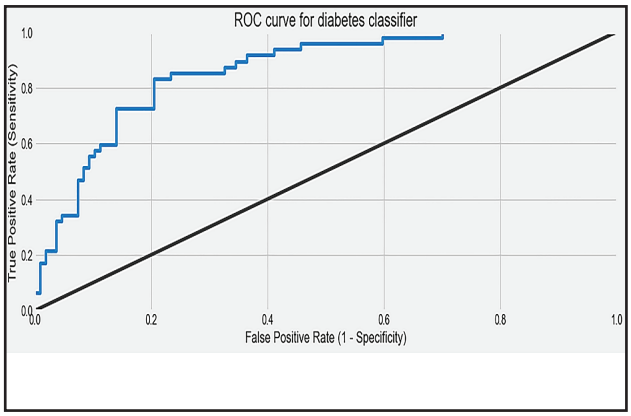

Figure 8. The Receiver Open Characteristic Curve (ROC).

The model's factor significance scores might give information about the model. A prediction model that is fitted to the dataset calculates the most important scores. When creating a prediction, checking the importance score offers insight into that specific model and the most essential and least significant aspects of the model. For models that support it, this is a sort of model interpretation that may be done. A prediction model's importance can be utilized to enhance it. This may be accomplished by utilizing significance scores to determine which elements to eliminate (those with the lowest scores) and which factors to maintain (those with the highest scores). As displayed in Figure 9, glucose has the highest score of 0.34, BMI of 0.18, age about 0.17, Diabetes Pedigree Function has a score of 0.09, the number of pregnancies and insulin have 0.06, and blood pressure and skin thickness are a score of 0.05 and 0.045; respectively.

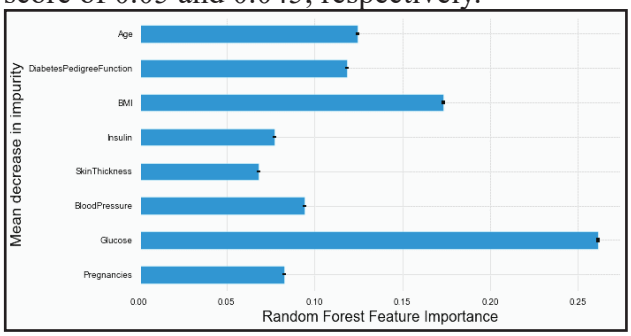

Figure 9. The Random Forest feature's importance for the classification decision.

The results could be limited to the period in which the data was gathered (between the 1960s and 1980s). A urinalysis, abdominal fat, LDL, HDL, cholesterol, inter-pregnancy interval, and

HbA1c test which indicates the average level of blood sugar over the past three months (33), are now utilized to diagnose diabetes mellitus. It is also observed that the Skin thickness values do not reflect the BMI values, however, the skin became steadily thicker with increasing BMI (34). Secondly, the Deep Cascade Forests technique is utilized with the same hyperparameters to improve the proposed predictive models. In this experiment, the achieved accuracy is about 98.6%. All predictive models were evaluated in the training set using a k-fold cross-validation approach.

The key benefit of this evaluation method is that it increases the amount of data available for training the models by allowing all data instances to be utilized for both training and validation in various iterations. It also provides reliable assessments of the performance of prediction models for data that has not yet been observed. In addition, Ensemble Learning techniques are based on the bootstrap method which is a resampling approach commonly utilized to estimate statistics on a population as well as validate the model. These methods develop a predictive model in each bootstrap sample and estimate measures of predictive performance in each bootstrap sample. The bootstrap models are then applied to the original dataset to validate the model and determine the predictive measure in the original data.

Based on applied techniques, it is highlighted that each one reached different associated factors to gestational diabetes. Concerning the Correlation heatmap, the associated factors with diabetes include Glucose, BMI, Age, Diabetes Pedigree Function, number of Pregnancies, and Insulin. While the Stepwise technique revealed that Glucose, Blood Pressure, Diabetes Pedigree Function, Insulin, and BMI are the most influential factors. However, the Random Forest technique selected Glucose, BMI, Diabetes Pedigree Function, Age, and Blood Pressure as important factors. Moreover, the Multivariate analysis and Generalized Linear Model revealed that important factors include, BMI > 25, Blood Pressure > 80, Diabetes Pedigree Function > 0.33, Glucose > 150, and Insulin > 400. However, MCA reached a model with including factors, Glucose > 150, Age > 35, Skin Thickness > 40, and Pregnancies > 6. In addition to these factors, the HCA revealed that Blood Pressure > 80 is an important risk factor. All obtained results are gathered in the scheme in Figure 10.

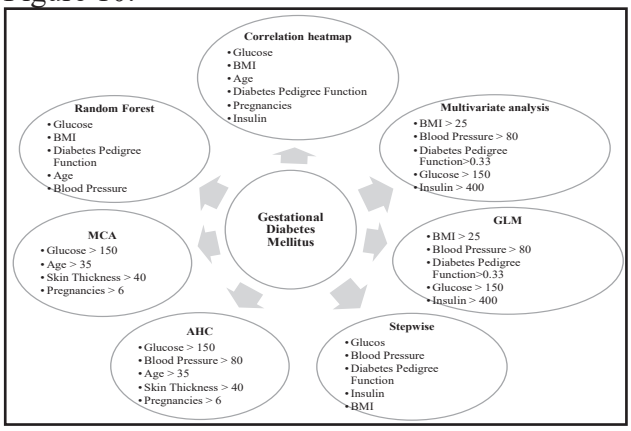

Figure 10. Potential factors of developed predictive models.

Discussion

In this work, gestational diabetes complications and their associated risk factors are discussed, in the studied population of patients of Pima Indian descent. A high prevalence of diabetes is reported due to several factors, including the number of Pregnancies, Age, BMI, Blood Pressure, Skin Thickness, Insulin, Glucose, and Diabetes Pedigree Function. The predictive models were developed by several statistical analyses and Machine Learning techniques on the PIMA Indian dataset for diabetes prevention. Based on the Correlation heatmap, associated factors with diabetes include Glucose, BMI, Age, Diabetes Pedigree Function, number of Pregnancies, and Insulin. However, the Stepwise technique revealed that Glucose, Blood Pressure, Diabetes Pedigree Function, Insulin, and BMI are the most influential factors. While the Random Forest technique selected Glucose, BMI, Diabetes Pedigree Function, Age, and Blood Pressure as important factors.

The Multivariate analysis and Generalized Linear Model revealed that important factors include, BMI > 25, Blood Pressure > 80, Diabetes Pedigree Function > 0.33, Glucose > 150, and Insulin > 400. However, MCA reached a model with including factors, Glucose > 150, Age > 35, Skin Thickness > 40, and Pregnancies > 6. In

addition to these factors, the HCA revealed that Blood Pressure > 80 is an important risk factor. To validate and compare the predictive models, an accuracy metric is performed. Cascade Forests has proven its effectiveness with a recognition rate of 98.58%. Random Forest achieved an accuracy of 89%. Compared to other classifiers, the accuracy is about 69% for SVM, for Logistic Regression, it is 78%, 78% also obtained for the Decision Tree, and when using K-NN, the reached accuracy is 72%. The development of these predictive models will make it possible to consider curative and preventive medico-health actions, in a prospective vision to gradually control these types of risks. Consistently to several studies, weight gain during pregnancy has a relation to GDM and hypertension which results in negative outcomes (35,36). Among 15 to 20% of pregnant women were susceptible to GDM or diabetes type 2, specifying that the rate of gestational diabetes increased in recent years (37). Advanced maternal age (38,39) and high BMI (37) are all considered factors favoring the occurrence of GDM or diabetes type 2. Several works reported an increased risk of GDM among overweight women (40,41).

Physical activity-based therapies lowered the risk of GDM by a significant amount (42,43). Although GDM was more frequent in women with advanced age and higher parity, the sibship design was suited for GDM in the second pregnancy or later pregnancies (44,45). Furthermore, GDM complicates 3 to 5% of pregnancies in the United States, and the prevalence is rising. Type 2 diabetes, high blood pressure, and cardiovascular disease are all substantial risks for women with GDM (46). Many pregnant women who were put on insulin continued to have trouble controlling their blood sugar, and they were concerned about the short- and long-term health implications of the drug on themselves and their newborns (47).

The number of pregnancies is considered an associated factor with GDM consistently in several research studies (33,48). Blood pressure in its turn is identified as a strong significant risk factor with GDM as identified in many works effectuated on this subject (49,50). Factors presented have already been found as risk factors in various studies and have been related to diabetes in the investigated research. Thus, in this study, these factors formed a predictive model of GDM that should help to prevent its complications while also protecting women and babies. In addition, proposing a guide and code of practice that can be essential as a decision-making tool for healthcare practitioners, by recommending to expectant mothers at high risk of GDM a satisfactory pre- and post-conceptual follow-up to minimize the occurrence of this risk.

The limitations that influence the quality of the result obtained are several, including:

- The Machine Learning techniques: some techniques are much better than others for a particular task, and using Deep Learning techniques may result in better accuracy and improved prediction.
- Data quality: missing data affects the prediction and minimizes the amount of data.
- The amount of training data (small data with 768 patients): from a few examples, it is difficult to attain a good predictive model.
- The number of features utilized to classify patients: just eight features (Number of pregnancies, Age, Glucose, Blood pressure, Skin Thickness, Insulin, BMI, and Diabetes Pedigree Function), the need for other medical tests is indispensable.

Conclusions

From a course life perspective, fertility, family planning, and pregnancy considerations are of increasing importance to any pregnant woman. Even though most pregnancies proceed without a hitch or incident and end up being perfect, some complications do occur. In this study, the prevention of gestational diabetes mellitus is treated in 768 women from the Pima Indian dataset. Factors discussed (number of Pregnancies, Age, BMI, Blood Pressure, Skin Thickness, Insulin, Glucose, and Diabetes Pedigree Function) are identified as risk factors in several studies, and they are demonstrated to be associated with diabetes in the studied population. Thus, these factors constitute a predictive model of GDM which may help to prevent its complications and at the same time protect women and babies. However, this

study proves that the importance and influence of factors are variable from one technique to another. Within these predictive models, it is possible to suggest a guide of practice that can be crucial as a decision support tool for public health professionals, by suggesting to expectant mothers at high risk of developing gestational diabetes an adequate pre- and post-conceptual follow-up to eliminate the occurrence of this risk (GDM) or to minimize its bearing.

The Deep Cascade Forests technique is carried out proving its effectiveness with an accuracy of 98.58%. The Random Forests technique achieved a recognition rate of 89% compared to other developed machine learning models. In addition to the K-Nearest Neighbor, SVM, and Logistic Regression methods. Besides the efficient statistical analysis techniques such as univariate and multivariate analysis, the Generalized Linear Model, Downtown Stepwise, HCA, and MCA, Machine Learning techniques present a very powerful tool for building predictive models.

Considering the rich results obtained by the predictive model based on gathered data about patients, it should be able to make better predictions on how likely women could suffer the onset of gestational diabetes and estimate its prevalence, therefore acting appropriately and rapidly to assist them. Future work is suggested to study the influence of nutritional assessments (micronutrients, vitamins, and minerals) on the occurrence of gestational diabetes. In addition to the search for genetic parameters and markers responsible for the vulnerability of the population to different diseases.

Acknowledgments

The authors would like to take this opportunity to express their profound gratitude and deep regard for everyone who helped in this work. This paper and the research behind it would not have been possible without the exceptional support of colleagues with provided insight and expertise that greatly assisted the research and the technical assistance. The authors wish to extend their special thanks and greet everyone, who has supported the creation of this work.

Statements and Declarations

Funding: No funds, grants, or other support were received.

Conflict of Interest: The authors declare that they have no conflict of interest.

Ethical Approval: Not included. The data is publicly available. It belongs to the National Institute of Diabetes, Digestive, and Kidney Diseases. Publicly available in the Pima Indians Diabetes Database repository, via the link: https://www.kaggle.com/uciml/pima-indians-diabetes-database/data.

Consent to participate/for publication involvement statement: Not included.

Data availability: The datasets generated and/or analyzed during the current study are available in the Pima Indians Diabetes Database repository, https://www.kaggle.com/uciml/pima-indians-diabetes-database/data

Code availability Not available.

Authors' Contribution H. Zermane and A. Kalla led the study by being involved in data interpretation, performing the statistical analysis, coding, interpreting the results, conceiving, and designing the article. All authors provided intellectual content revisions for clarity and precision on the subject matter.

References

1. WHO. Diabetes [Internet]. 2021 [cited 2021 Oct 30]. Available from: https://www.who.int/news-room/fact-sheets/detail/diabetes
2. Kautzky-Willer A, Harreiter J, Winhofer-Stöckl Y, Bancher-Todesca D, Berger A, Repa A, et al. Gestational diabetes mellitus (Update 2019). Wiener Klinische Wochenschrift. 2019;131:91–102.
3. Kotzaeridi G, Blätter J, Eppel D, Rosicky I, Falcone V, Adamczyk G, et al. Recurrence of Gestational Diabetes Mellitus : To Assess Glucose Metabolism and Clinical Risk Factors at the Beginning of a Subsequent Pregnancy. Journal of Clinical Medicine. 2021;10(7494):1–10.
4. Schwartz N, Nachum Z, Green MS. Risk factors of gestational diabetes mellitus recurrence: a meta-analysis. Endocrine.

2016;53(3):662–71.
5. Utz B, Kolsteren P, De Brouwere V. Screening for gestational diabetes mellitus: Are guidelines from high-income settings applicable to poorer countries? Clinical Diabetes. 2015;33(3):152–8.
6. McIntyre HD, Catalano P, Zhang C, Desoye G, Mathiesen ER, Damm P. Gestational diabetes mellitus. Nature Reviews Disease Primers. 2019;5(1).
7. Levy A, Wiznitzer A, Holcberg G, Mazor M, Sheiner E. Family history of diabetes mellitus as an independent risk factor for macrosomia and cesarean delivery. Journal of Maternal-Fetal and Neonatal Medicine. 2010;23(2):148–52.
8. Mercaldo F, Nardone V, Santone A. Diabetes mellitus affected patients classification and diagnosis through machine learning techniques. Procedia Computer Science. 2017;112:2519–28.
9. Bagley SC, White H, Golomb BA. Logistic regression in the medical literature: Standards for use and reporting, with particular attention to one medical domain. Journal of Clinical Epidemiology. 2001;54(10):979–85.
10. LaValley MP. Logistic regression. Circulation. 2008;117(18):2395–9.
11. El Sanharawi M, Naudet F. Understanding logistic regression. Journal Francais d'Ophtalmologie. 2013;36(8):710–5.
12. Hosmer DW, Lemeshow S, Sturdivant RX. Applied Logistic Regression: Third Edition. Applied Logistic Regression: Third Edition. 2013. 1–510 p.
13. Saberioon M, Císař P, Labbé L, Souček P, Pelissier P, Kerneis T. Comparative performance analysis of support vector machine, random forest, logistic regression and k-nearest neighbours in rainbow trout (oncorhynchus mykiss) classification using image-based features. Sensors (Switzerland). 2018;18(4):1–15.
14. Gholipour K, Asghari-Jafarabadi M, Iezadi S, Jannati A, Keshavarz S. Modelling the prevalence of diabetes mellitus risk factors based on artificial neural network and multiple regression. Eastern Mediterranean health journal = La revue de sante de la Mediterranee orientale = al-Majallah al-sihhiyah li-sharq al-mutawassit. 2018;24(8):770–7.
15. Chatterjee S, Goyal D, Prakash A, Sharma J. Exploring healthcare/health-product ecommerce satisfaction: A text mining and machine learning application. Journal of Business Research. 2021;131(October):815–25.
16. Hui EGM. Learn R for Applied Statistics. Learn R for Applied Statistics. 2019.
17. Contreras I, Vehi J. Artificial intelligence for diabetes management and decision support: Literature review. Journal of Medical Internet Research. 2018;20(5):1–21.
18. Vapnik VN. Pattern Recognition-Statistical Learning Theory. Canada: Wiley; 1998. 1–760 p.
19. Vapnik VN. The Nature of Statistical Learning. Theory. 1995.
20. Zermane H, Kasmi R. Intelligent industrial process control based on fuzzy logic and machine learning. International Journal of Fuzzy System Applications. 2020;9(1):92–111.
21. Hsu CW, Lin CJ. A comparison of methods for multiclass support vector machines. IEEE Transactions on Neural Networks. 2002;13(2):415–25.
22. Kale R, Shitole S. Analysis of Crop disease detection with SVM, KNN and Random forest classification. Information Technology in Industry. 2021;9(1):364–72.
23. Rahab H, Zitouni A, Djoudi M. SIAAC: Sentiment Polarity Identification on Arabic Algerian Newspaper Comments. Advances in Intelligent Systems and Computing. 2018;662:139–49.
24. Houfani D, Slatnia S, Kazar O, Zerhouni N, Saouli H RI. Breast cancer classification using machine learning techniques: a comparative study. Medical Technologies Journal. 2020;4(2):535–44.
25. Elaziz MA, Hosny KM, Salah A, Darwish MM, Lu S, Sahlol AT. New machine learning method for image-based diagnosis of COVID-19. PLoS ONE. 2020;15(6):1–18.
26. Ko BC, Kim SH, Nam JY. X-ray image classification using random forests with local wavelet-based CS-local binary patterns. Journal of Digital Imaging. 2011;24(6):1141–51.

27. Dino HI, Abdulrazzaq MB. Facial Expression Classification Based on SVM, KNN and MLP Classifiers. 2019 International Conference on Advanced Science and Engineering, ICOASE 2019. 2019;70–5.
28. Dietterich TG. Experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. Machine Learning. 2000;40(2):139–57.
29. Zermane H, Drardja A. Development of an efficient cement production monitoring system based on the improved random forest algorithm. International Journal of Advanced Manufacturing Technology. 2022;120(3–4):1853–66.
30. Zermane A, Mohd Tohir MZ, Zermane H, Baharudin MR, Mohamed Yusoff H. Predicting fatal fall from heights accidents using random forest classification machine learning model. Safety Science. 2023;159(November 2022):106023.
31. Breiman L. Random forests. Machine Learning. 2001;45(1):5–32.
32. Zhou ZH, Feng J. Deep forest. National Science Review. 2019;6(1):74–86.
33. Plows JF, Stanley JL, Baker PN, Reynolds CM, Vickers MH. The pathophysiology of gestational diabetes mellitus. International Journal of Molecular Sciences. 2018;19(11):1–21.
34. Gibney MA, Arce CH, Byron KJ, Hirsch LJ. Skin and subcutaneous adipose layer thickness in adults with diabetes at sites used for insulin injections: Implications for needle length recommendations. Current Medical Research and Opinion. 2010;26(6):1519–30.
35. Gou BH, Guan HM, Bi YX, Ding BJ. Gestational diabetes: Weight gain during pregnancy and its relationship to pregnancy outcomes. Chinese Medical Journal. 2019;132(2):154–60.
36. Zheng W, Huang W, Liu C, Yan Q, Zhang L, Tian Z, et al. Weight gain after diagnosis of gestational diabetes mellitus and its association with adverse pregnancy outcomes: a cohort study. BMC Pregnancy and Childbirth. 2021;21(1):1–9.
37. Heude B, Thiébaugeorges O, Goua V, Forhan A, Kaminski M, Foliguet B, et al. Pre-pregnancy body mass index and weight gain during pregnancy: Relations with gestational diabetes and hypertension, and birth outcomes. Maternal and Child Health Journal. 2012;16(2):355–63.
38. Ben-David A, Glasser S, Schiff E, Zahav AS, Boyko V, Lerner-Geva L. Pregnancy and Birth Outcomes Among Primiparae at Very Advanced Maternal Age: At What Price? Maternal and Child Health Journal. 2016;20(4):833–42.
39. Fuchs F, Monet B, Ducruet T, Chaillet N, Audibert F. Effect of maternal age on the risk of preterm birth: A large cohort study. Obstetrical and Gynecological Survey. 2018;13(1):1–10.
40. Andreasen KR, Andersen ML, Schantz AL. Obesity and pregnancy. Acta Obstet Gynecol Scand. 2004;83(11):1022--1029.
41. Cedergren MI. Maternal morbid obesity and the risk of adverse pregnancy outcome. Obstetrics and gynecology. 2004;103(2):219–24.
42. Peters TM, Brazeau AS. Exercise in Pregnant Women with Diabetes. Current Diabetes Reports. 2019;19(9).
43. Leppänen M, Aittasalo M, Raitanen J, Kinnunen TI, Kujala UM, Luoto R. Physical activity during pregnancy: predictors of change, perceived support and barriers among women at increased risk of gestational diabetes. Maternal and child health journal. 2014;18(9):2158–66.
44. Yu Y, Arah OA, Liew Z, Cnattingius S, Olsen J, Sørensen HT, et al. Maternal diabetes during pregnancy and early onset of cardiovascular disease in offspring: Population based cohort study with 40 years of follow-up. The BMJ. 2019;367(Cvd):1–4.
45. Davenport MH, Ruchat SM, Poitras VJ, Jaramillo Garcia A, Gray CE, Barrowman N, et al. Prenatal exercise for the prevention of gestational diabetes mellitus and hypertensive disorders of pregnancy: A systematic review and meta-analysis. British Journal of Sports Medicine. 2018;52(21):1367–75.
46. Kalla A, Loucif L, Yahia M. Miscarriage Risk Factors for Pregnant Women: A Cohort Study in Eastern Algeria's Population. The Journal of Obstetrics and Gynecology of

India. 2022 Aug;72(Suppl 1):109-120.
47. Figueroa Gray M, Hsu C, Kiel L, Dublin S. “It’s a Very Big Burden on Me”: Women’s Experiences Using Insulin for Gestational Diabetes. Maternal and Child Health Journal. 2017;21(8):1678–85.
48. Liu B, Song L, Zhang L, Wang L, Wu M, Xu S, et al. Higher numbers of pregnancies associated with an increased prevalence of gestational diabetes mellitus: Results from the healthy baby cohort study. Journal of Epidemiology. 2020;30(5):208–12.
49. Yan B, Yu Y, Lin M, Li Z, Wang L, Huang P, et al. High, but stable, trend in the prevalence of gestational diabetes mellitus: A population-based study in Xiamen, China. Journal of Diabetes Investigation. 2019;10(5):1358–64.
50. Sibai BM, Ross MG. Hypertension in gestational diabetes mellitus: Pathophysiology and long-term consequences. Journal of Maternal-Fetal and Neonatal Medicine. 2010;23(3):229–33.