Predicting Diabetes Onset In Female Patients Using Machine Learning: A Logistic Regression Approach

A. Poompavai[1], A. Poongothai[2]

[1]Assistant Professor, [2]Research Scholar

Department of Statistics, SDNB Vaishnav College, Chrompet, Chennai

Department of Statistics, Annamalai University, Chidambaram.

Poompavai.a@sdnbvc.edu.in, poongothai.a93@gmail.com

ABSTRACT

Artificial Intelligence and Machine Learning are increasingly being used in healthcare for early diabetes detection and management. A dataset of 768 records of female patients, each characterized by eight health-related attributes, is used to predict the onset of diabetes. The dataset includes columns such as pregnancy, glucose, blood pressure, skin thickness, insulin, BMI, diabetes pedigree function, age, and outcome. The dataset is adapted from the National Institute of Diabetes and Digestive and Kidney diabetes outcome, helping detect gestational diabetes, blood pressure, and diabetes, which can lead to Diseases (NIDDK) and is widely used in machine learning research on healthcare and medical diagnostics. The dataset uses bar graphs and histograms to visually represent categorical variables of serious health issues. The logit regression model predicts significant differences between outcomes and Skin Thickness, Insulin, and Age, but no significant differences between outcomes and Intercept, Pregnancies, Glucose, BP, BMI, and Diabetes Pedigree Function. The binary logistic model has an AUC of 0.84, accuracy of 75.78%, no information rate of 0.8203, kappa-value of 0.43, sensitivity of 0.8913, specificity of 0.7286, precision of 0.4184, recall of 0.8913, and F1 Score of 0.5694.

Keywords: **Diabetes, Explanatory Data Analysis (EDA), Logit Regression, Area Under ROC Curve (AUC), Confusion Matrix (CM), Model Statistics.**

I. INTRODUCTION

Diabetes is a metabolic disorder characterized by high blood sugar levels, which can damage vital organs and cause other health issues. There are two types: Type 1 Diabetes Mellitus, more common in children, and Type 2 Diabetes Mellitus, attributed to lifestyle factors such as obesity, low physical activity, an unhealthy diet, and excessive sugar consumption. Diabetes is caused by the breakdown of food into glucose, which the pancreas supplies to the body. When the pancreas fails to produce enough insulin or body cells do not receive glucose, diabetes results in excess glucose in the blood. Symptoms include fatigue, irritation, stress, tiredness, frequent urination, headache, loss of strength and stamina, weight loss, and increased appetite. There are two types of blood sugar levels: fasting blood sugar level (100 mg/dL before eating food) and postprandial glucose level (under 140 mg/dL two hours after meals). Diabetes can cause severe health consequences, including damage to vital body organs, cardiovascular and skin diseases, and increased risk of heart attacks. Artificial Intelligence and Machine Learning have made significant strides in predicting health emergencies, disease populations, and immune response. Despite skepticism about their practical application in healthcare, their inclusion is increasing rapidly. These technologies are promising for early diabetes detection and management, using extensive datasets to identify patterns and risk factors associated with the disease.

II. REVIEW OF LITERATURE

The study by Taghreed M Jawa (2022) examines the impact of home quarantine on psychological stability among 846 residents of Makkah during the COVID-19 pandemic. The research used a multiple logistic regression model to analyze the relationship between factors such as education level and psychological disorders during home quarantine. The results showed that a high percentage of respondents felt psychological stability during the quarantine period. The study also found that home quarantine positively impacted health, self-development, and family closeness. However, private employees or unemployed individuals experienced more psychological disorders and decreased income due to home quarantine.

Lee E.J., et al., (2017) highlight the increasing interest in Artificial Intelligence (AI) in stroke medicine, particularly in improving diagnosis accuracy and patient care. AI techniques have shown promising results in deciphering data from stroke imaging, and in the future, they may play a pivotal role in determining therapeutic methods and predicting prognosis for stroke patients in an individualized manner. This review explores the technical principles, clinical application, and future perspectives of AI in stroke imaging.

Joanne Peng et al.'s (2002) article provides guidelines for using logistic regression techniques in articles. It discusses the use of tables, figures, and charts to assess results and verify assumptions. The article also provides recommendations for reporting formats and minimum observation-to-predictor ratio. The authors evaluated 8 articles in The Journal of Educational Research between 1990 and 2000, finding that all studies met or exceeded recommended criteria.

Khatimya Kudabayeva et al. (2024) conducted a study on LADA type diabetes, a subtype of diabetes with a later onset. The study analyzed 672 publications from 1994 to 2024, focusing on patients aged 18 and older. The analysis revealed an upward trend in annual publications, with China leading with 101 papers published. The study also highlighted the importance of international collaboration and the contributions of leading countries and institutions in shaping our understanding of this complex subtype of diabetes. The study highlights the importance of international collaboration in diabetes research.

III. DATABASE

The dataset is a collection of 768 records of female patients, each characterized by eight health-related attributes, aimed at predicting the onset of diabetes. The outcome variable indicates whether the patient has diabetes or not. The columns include pregnancy, glucose, blood pressure, skin thickness, insulin, BMI, diabetes pedigree function, age, and outcome. The dataset can be used for building classification models, exploratory data analysis, feature engineering, and feature selection for healthcare-related datasets. The dataset is adapted from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) and is widely used in machine learning research on healthcare and medical diagnostics.

IV. METHODOLOGY

BAR CHART

Bar charts, also known as bar graphs, are a popular and effective method for visually representing categorical data in a structured manner. They are pictorial representations of data with rectangular bars with heights or lengths proportional to the values they represent, containing numerical values of variables representing length or height.

HISTOGRAM AND DENSITY PLOT

Histograms and density plots are statistical tools used to visualize data distribution. Histograms group values into continuous ranges, with each bar representing the number of values in that range. They display the frequency of each value in a dataset, while density plots show the proportion of data in each range. Both methods are useful for understanding data patterns and can be used for various data analysis.

BOXPLOT

Boxplots are a useful tool for comparing multiple data sets in a standard format, displaying the centre and spread of a unimodal data set along a numeric variable using five summary values.

CORRELATION HEATMAP

A correlation heat map is a color-coded matrix that displays the correlation between multiple variables, resembling a colour chart. Each variable is represented by a row and column, with each cell's colour representing the strength and direction of the correlation, with darker colours indicating stronger ones.

LOGISTIC REGRESSION

Logistic regression is a classification algorithm used to predict binary outcomes, such as success or failure, using the logit function as a link function in a binomial distribution. It finds the best fit between predictor and target variables to predict the probability of the target variable belonging to a labeled class. This statistical modeling technique is widely used in various industries, including marketing, finance, social sciences, and medical research. The sigmoid function, also known as logistic function, describes the correlation between predictor variables and the likelihood of the binary outcome.

Logistics regression is a generalized linear model used to predict qualitative responses, with a value of y ranging from 0 to 1, represented by the equation.

$$\text{Odds} = \frac{P}{1-P}$$

The odds ratio, a key representation of logistic regression coefficients, represents the probability of success compared to failure, ranging from 0 to infinity.

$$\log(\text{Odds}) = \log\left(\frac{P}{1-P}\right)$$

We must select a link function that is optimal for the binomial distribution, as it is a dependent variable.

$$\text{logit}(P) = \log\left(\frac{P}{1-P}\right) = b0+b1X1+b2X2+ \ldots. +bkXk$$

The logit function, also known as a log of odds, is a linear function that maximizes the likelihood of observing sample values. It must be linearly related to the independent variables, as assumed in the OLS assumption. Variables like b0, b1, and b2 are unknown and must be estimated using available training data. In a logistic regression model, multiplying b1 by one unit changes the logit by b0, with P changing based on the multiplied value.

Classification evaluation metrics are quantitative measures used to assess the performance and accuracy of a classification model. They provide insights into how well a model can classify instances into predefined classes or categories. The confusion matrix is a useful tool for evaluating classification problems with multiple classes, while accuracy measures the overall correctness of a model's predictions. Precision focuses on the correctness of positive predictions, especially in scenarios where false positives have significant consequences. Recall measures the model's ability to correctly identify positive instances. The F1 score balances precision and recall, while specificity measures the model's ability to correctly identify negative instances.

Accuracy = (TP+FP) / (P+N)

Precision = TP / (TP+FP)

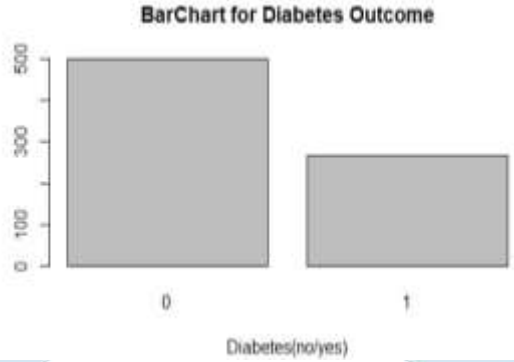
Recall (Sensitivity or True Positive Rate) = TP / (TP+FN)

F1 Score = (2 X Precision X Recall) / (Precision + Recall)

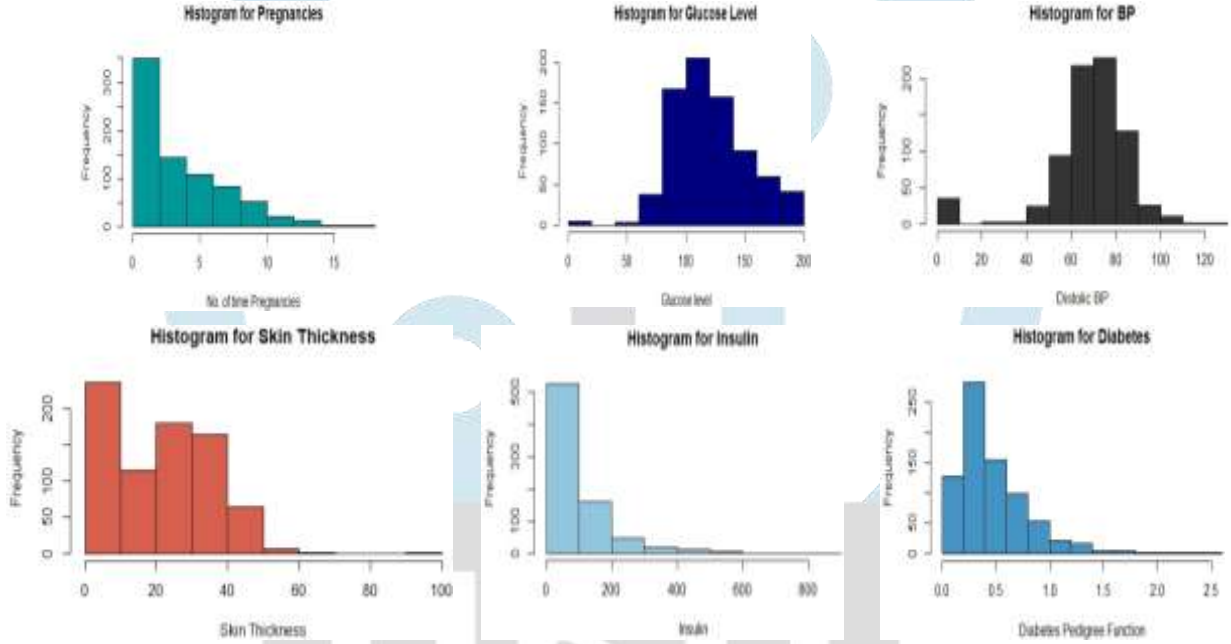
Specificity (True Negative Rate) = TN / (TN+FP)

Kappa Score = (observed – expected) / (100% - expected)

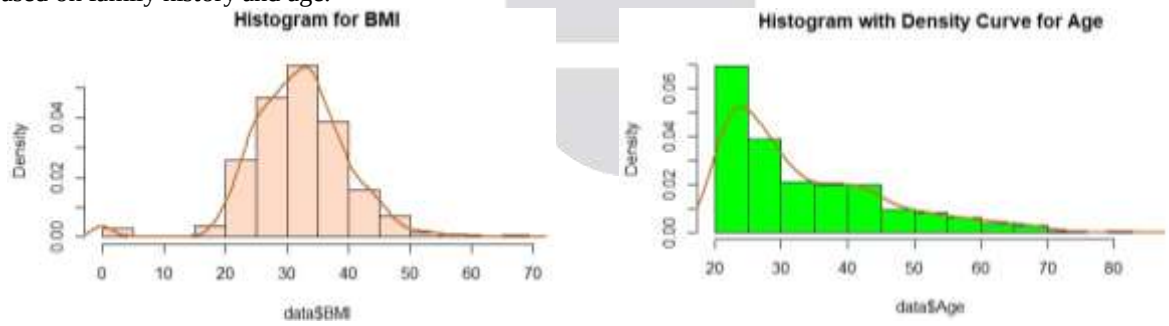
V. RESULT AND CONCLUSION



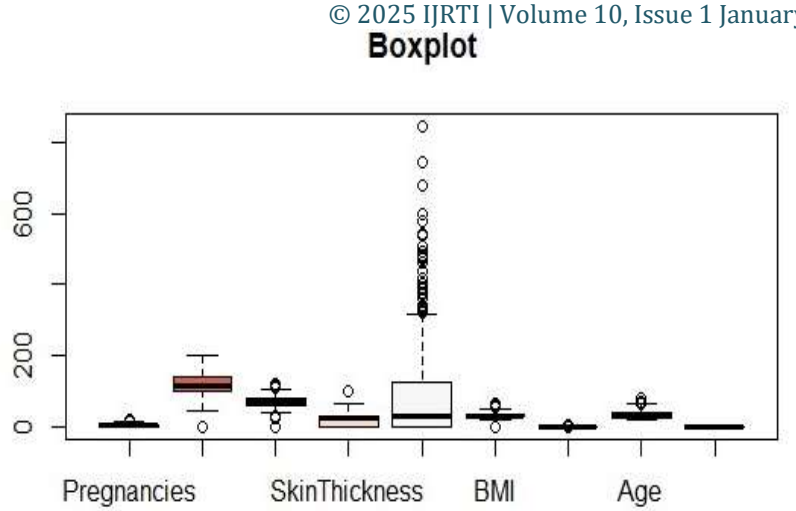
The bar graph visually represents categorical variables of diabetes outcome, with values of 500 and 268, using pictorial rectangular bars with heights proportional to their respective values.



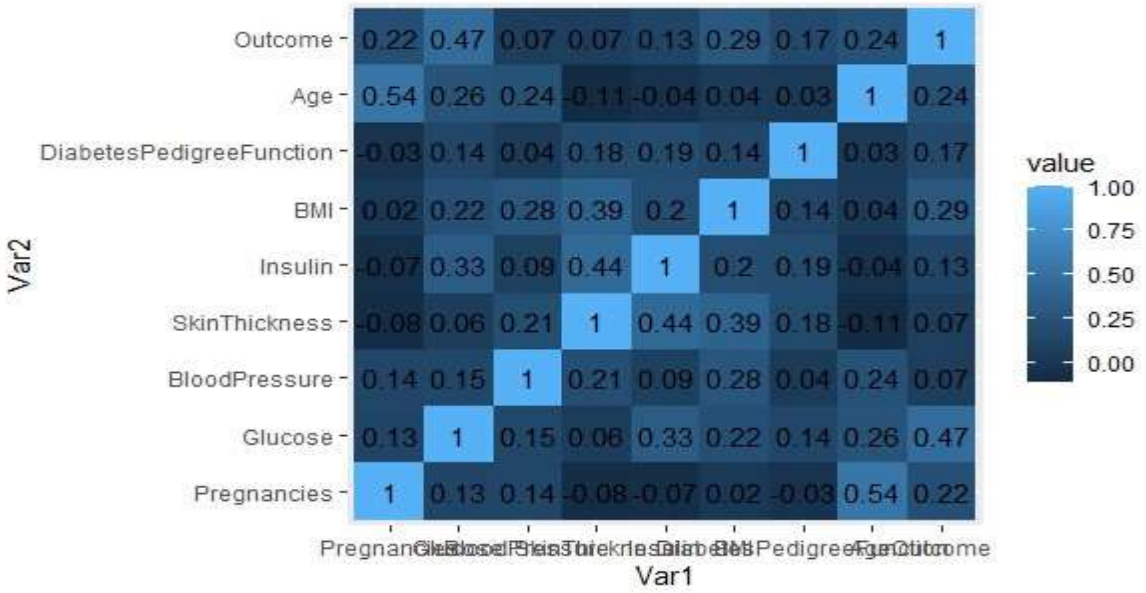
A histogram is a statistical tool used to display the frequency of female pregnancies, glucose tolerance tests, blood pressure, and skinfold thickness measurements. It helps detect gestational diabetes, blood pressure, and diabetes, which can lead to serious health issues. The glucose tolerance test measures the body's response to sugar and starch after a meal, while the diabetes probability function calculates the likelihood of diabetes based on family history and age. Normal blood pressure ranges from 120-129 to 80-84 mm Hg. The histogram also shows the 'Diabetes Pedigree Function' function, which calculates diabetes probability based on family history and age.



The histogram and density plot are used to visualize the BMI and Age data in the diabetes dataset, providing a comprehensive understanding of data patterns and facilitating data analysis.



Boxplots are a useful tool for visualizing data sets by displaying important information in a simple, standard format. They use five summary values: minimum, first quartile, median, third quartile, and maximum. The box diagram begins with the first quartile and ends with the third quartile, with the median as a line dividing it. Lines extend from the first quartile to the minimum and from the third quartile to the maximum. 25% of the data values fall between these quartiles.



A correlation heat map is a color-coded matrix that displays the correlation between multiple variables, resembling a colour chart. It represents each variable by row and column, with lighter colours indicating stronger correlations. This visualization helps show the relationship between different dataset variables.

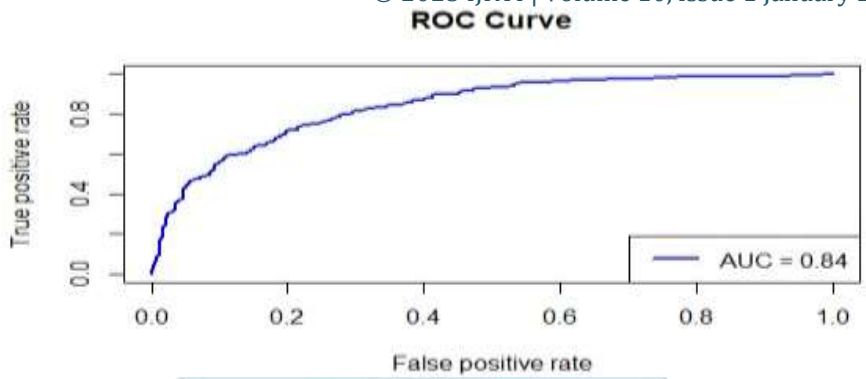
Table 1 Logistic Regression

Variables	Estimates	Std. Error	Z value	P value
(Intercept)	-8.4046964	0.7166359	-11.728	< 2e-16
Pregnancies	0.1231823	0.0320776	3.840	0.000123
Glucose	0.0351637	0.0037087	9.481	< 2e-16
Blood Pressure	-0.0132955	0.0052336	-2.540	0.011072
Skin Thickness	0.0006190	0.0068994	0.090	0.928515
Insulin	-0.0011917	0.0009012	-1.322	0.186065
BMI	0.0897010	0.0150876	5.945	2.76e-09
Diabetes Pedigree Function	0.9451797	0.2991475	3.160	0.001580
Age	0.0148690	0.0093348	1.593	0.111192

The logit regression line is

Y (Outcome) = -8.4047 + 0.1232 Pregnancies + 0.0352 Glucose – 0.0133 BP + 0.0006 Skin Thickness – 0.0012 Insulin + 0.090 BMI + 0.9452 Diabetes Pedigree Function + 0.015 Age.

There is a significant difference between the outcome and Skin Thickness, Insulin & Age. There is no significant difference between the Outcome and Intercept, Pregnancies, Glucose, BP, BMI & Diabetes Pedigree Function. Null deviance (993.48 on 767 d.f) is the value when only one variable is present in an equation, while residual deviance (723.45 on 759 d.f) takes all variables into account. A model is considered good if the difference between these two values is significant. A larger difference indicates a good fit, and a larger difference indicates a better model. The Akaike Information Criterion (AIC) value is 741.45 and the number of Fisher scoring iterations is 5. The area under the ROC curve (AUC) is the probability that a model will rank the positive higher than the negative, and the binary logistic model has an AUC of 0.84 and it is shown in the below figure.



The confusion Matrix for logistic Regression is

Table 2. Confusion Matrix

OUTCOMES	0	1
0	153	5
1	57	41

Table 3. Classification Metrics

Classification Evaluation Metrics	Value
Accuracy	75.78%
95% CI	0.7006 – 0.809
No Information Rate	0.8203
p-value (Acc > NIR)	0.9953
Kappa	0.43
Mcnemar's Test p-value	9.356e-11
Sensitivity	0.8913
Specificity	0.7286
Pos Pred Value	0.4184
Neg Pred Value	0.9684
Precision	0.4184
Recall	0.8913
F1	0.5694
Prevalance	0.1797
Detection Rate	0.1602
Detection Prevalence	0.3828
Balanced Accuracy	0.8099

The model has an accuracy of 75.78%, a no information rate of 0.8203, a kappa-value of 0.43, a sensitivity of 0.8913, a specificity of 0.7286, a precision of 0.4184, a recall of 0.8913 and F1 Score of 0,5694.

VI. CONCLUSION

Artificial Intelligence and Machine Learning have made significant strides in predicting health emergencies, disease populations, and immune response. They are increasingly being used in healthcare for early diabetes detection and management. A dataset of 768 records of female patients, each characterized by eight health-related attributes, is used to predict the onset of diabetes. The outcome variable indicates whether the patient has diabetes or not. The dataset includes columns such as pregnancy, glucose, blood pressure, skin thickness, insulin, BMI, diabetes pedigree function, age, and outcome. The dataset is adapted from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) and is widely used in machine learning research on healthcare and medical diagnostics. The bar graph visually represents categorical variables of diabetes outcome, while a histogram displays the frequency of female pregnancies, glucose tolerance tests, blood pressure, and skinfold thickness measurements. This helps detect gestational diabetes, blood pressure, and diabetes, which can lead to serious health issues. Boxplots are useful tools for visualizing data sets by displaying important information in a simple, standard format. They use five summary values: minimum, first quartile, median, third quartile, and maximum. A correlation heat map displays the correlation between multiple variables, resembling a color chart. There is a significant difference between the outcome and Skin Thickness, Insulin & Age, but no significant difference between the Outcome and Intercept, Pregnancies, Glucose, BP, BMI & Diabetes Pedigree Function. A model is considered good if the difference between these two values is significant. The Akaike Information Criterion (AIC) value is 741.45, and the area under the ROC curve (AUC) is the probability that a model will rank the positive higher than the negative.

VII. REFERENCES

1. Lee E.J., Kim Y.H., Kim N., Kang D.W. Deep into the brain: Artificial intelligence in stroke imaging. J. Stroke. 2017; 19(3):277–285.
2. Miotto R., Wang F., Wang S., Jiang X., Dudley J.T. Deep learning for healthcare: Review, opportunities and challenges. Brief. Bioinform. 2018;19(6):1236–1246.

3. Alloghani M., Al-Jumeily D., Mustafina J., Hussain A., Aljaaf A.J. A systematic review on supervised and unsupervised machine learning algorithms for data science. In: Berry M., Mohamed A., Yap B., editors. Supervised and unsupervised learning for data science. Springer, US: 2020. pp. 3–21.
4. Sandhu, T. H. (2018). Machine learning and natural language processing—A review. International Journal of Advanced Research in Computer Science, 9(2), 582–584.
5. Chao-Ying Joanne Peng, Kuk Lida Lee and Gary M. Ingersoll (2002) An Introduction to Logistic Regression Analysis and Reporting, The Journal of Educational Research, Volume 96, Issue !, PP: 3-14.
6. Cabrera, A. F. (1994). Logistic regression analysis in higher education: An applied perspective. Higher Education: Handbook of Theory and Research, Vol. 10, 225–256.
7. Chuang, H. L. (1997). High school youth's dropout and re-enrollment behavior. Economics of Education Review, 16(2), 171–186.
8. Peng, C. Y., & So, T. S. H. (2002b). Logistic regression analysis and reporting: A primer. Understanding Statistics, 1(1), 31–70.
9. Machamer, A. M., & Gruber, E. (1998). Secondary school, family, and educational risk: Comparing American Indian adolescents and their peers. The Journal of Educational Research, 91(6), 357–369.
10. Agnieszka Strzelecka, et al., (2020), Application of Logistic Regression models to assess household financial decisions regarding debt, 24th International Conference on Knowledge Based and Intelligent Information & Engineering Systems, Procedia Computer Science, Volume 176, PP: 3418-3427.
11. Taghreed M Jawa (2022), Logistic Regression Analysis for studying the impact of home quarantine on psychological health during COVID-19 in Saudi Arbia, National Library of Medicine – Alexandria Engineering Journal, Volume 61, Issue 10, PP:7995-8005.
12. Namasudra S., Dhamodharavadhani S., Rathipriya R. Nonlinear neural network based forecasting model for predicting COVID-19 cases. Neural Process. Lett. 2021:1–21.
13. Mahmoudi M.R., Heydari M.H., Qasem S.N., Mosavi A., Band S.S. Principal component analysis to study the relations between the spread rates of COVID-19 in high risks countries. Alexandria Eng. J. 2021;60(1):457–464.
14. Khatimya Kudabayeva et al., (2024), Global Trends in LADA Type Diabetes Research: A Bibliometric Analysis of Publications from Web of Science and Scopus, 1994-2024, Journal of Diabetes Research, Volume 2024, PP: 11 Pages.
15. Q. Wang, N. Li, and Q. Zhu, "Research on the developmenttrend of global diabetes health management based on patentbibliometrics," Chinese Journal of Health Management,vol. 16, 2022.
16. B. Zhang, Z. Jin, T. Zhai et al., "Global research trends on thelinks between the gut microbiota and diabetes between 2001and 2021: a bibliometrics and visualized study," Frontiers inMicrobiology, vol. 13, 2022.
17. T. Tuomi, Å. L. Carlsson, H. Li et al., "Clinical and geneticcharacteristics of type 2 diabetes with and without GAD anti-bodies," Diabetes, vol. 48, no. 1, pp. 150–157, 1999.
18. S. Carlsson, "Etiology and pathogenesis of latent autoimmunediabetes in adults (LADA) compared to type 2 diabetes," Fron-tiers in Physiology, vol. 10, 2019.
19. M. O. Gore, D. K. McGuire, I. Lingvay, and J. Rosenstock,"Predicting cardiovascular risk in type 2 diabetes: the hetero-geneity challenges," Current Cardiology Reports, vol. 17,no. 7, p. 607, 2015.
20. P. Pozzilli and U. Di Mario, "Autoimmune diabetes not requir-ing insulin at diagnosis (latent autoimmune diabetes of theadult)," Diabetes Care, vol. 24, no. 8, pp. 1460–1467, 2001.