UDC 519.2
DOI: https://doi.org/10.17721/1812-5409.2026/1.31

Mykola PYCHTAR, PhD (Ped.), Assoc. Prof.
ORCID ID: 0000-0003-4567-8901
e-mail: nppihtar@gmail.com
National Pedagogical Dragomanov University, Kyiv, Ukraine

Myroslava LYTVYN, Student
ORCID ID: 0009-0006-6792-6053
e-mail: myroslava.lytvyn.25@kse.org.ua
Kyiv School of Economics, Kyiv, Ukraine

STATISTICAL ANALYSIS OF FACTORS AFFECTING THE INCIDENCE OF DIABETES MELLITUS

The seriousness of the problem of diabetes mellitus is explained by the scale of its prevalence. In 2021, approximately 537 million people worldwide were living with diabetes, which accounts for about 6% of the global population. Therefore, the study and analysis of data on the dynamics of diabetes incidence are highly relevant, both for forecasting future trends and for supporting effective government regulation and prevention of the disease's spread.

The aim of this article is to conduct a statistical analysis of diabetes prevalence, to develop models of its dynamics for the city of Kyiv, and to analyze the influence of factors among women, such as the number of pregnancies, blood glucose levels, blood pressure, and body mass index (BMI), on the development of diabetes. Based on these variables, classification methods are applied to build predictive models using the R programming language.

K e y w o r d s: ***time series, diabetes mellitus, forecasting and modeling, k-NN, Random Forest, logistic regression, algorithm accuracy.***

AMS 2020 classification: 60G10, 60G15, 91-10.

Introduction

Diabetes is a metabolic disorder that affects many people worldwide. Every year, the incidence rates are rising at an alarming rate. If treatment is delayed, diabetes-related complications in many vital organs of the body can lead to a fatal outcome. It is very important to detect the disease at an early stage, which can halt its progression before complications arise. The International Diabetes Federation (IDF) has stated that currently, over 537 million people worldwide have diabetes. Over the last few years, the number of people with diabetes has risen sharply, making it a global threat. Unfortunately, diabetes now consistently ranks third among the leading causes of death.

Because early detection is vital to halting the disease's progression before fatal complications arise, researchers have increasingly turned to machine learning as a viable solution for automated prediction and diagnosis. Various robust frameworks have been proposed to address the complexities of medical data. The use of data analysis and machine learning methods can significantly improve the diagnosis of diabetes and help to take timely measures for its prevention and treatment. Due to the relevance of this direction, machine learning methods have been actively developed in recent years (Aguilera-Venegas et al., 2023; Kangra, & Singh, 2023; Olisah, Smith, & Smith, 2022; Orlova et al., 2026). Works in this area demonstrate the high efficiency of the classification models for various data sets. Recent advancements also include specialized deep learning architectures like EchoceptionNet, which captures both temporal and spatial feature interactions, and the comparative use of Decision Tree Classifiers and Artificial Neural Networks (ANN), with the former often demonstrating superior accuracy in identifying diabetic conditions based on biochemical indicators like HbA1c (Iftikhar et al., 2025). In (Livinska, Skrypnyk, & Galán-García, 2024), the accuracies of different machine learning algorithms were compared for diagnosing diabetes on binary data that do not include laboratory tests, and can be obtained by patient questionnaires.

This work examined data on people with diabetes in Kyiv and conducted a descriptive analysis to calculate key statistical characteristics; it also investigated two models based on ARIMA processes and linear regression for the percentage of children with diabetes. Using these, forecasts were constructed for the next 5 years. All this is described in more detail in the first chapter. The second chapter is devoted to the analysis of the diabetes incidence dataset in women, along with 9 other variables – factors that will be the focus of further work. The distribution of variables was analysed using histograms and box-and-whisker plots, and the relationships between them were investigated using a correlation matrix. The Student's t-test was used to compare the mean values of variables in the two groups (healthy, diabetic). Three machine learning methods were applied to classify patients into groups: k-nearest neighbours, logistic regression, and random forest. To compare these methods, the following quality measures were calculated: accuracy, sensitivity, specificity, and other indicators.

1. Statistical analysis and forecasting of the incidence of diabetes mellitus in Kyiv

This section is devoted to the analysis of data on the incidence of diabetes in Kyiv. Based on the data from the website (IDF, 2024). The data is obtained for 1995-2016 (no data for 2014, due to the annexation of Crimea and the east of the country. Since 2017, data for all regions have not been available, and their collection has not been continued.

Autocorrelation, also known as serial correlation, is the relationship between a signal (time series) and itself at different points in time. Informally, autocorrelation involves comparing two observations of the same signal at different times to assess how similar they are. Let's construct an autocorrelation function for the time series of the percentage of children with diabetes. To do this, our dataset must be fully known, so we impute the missing value for 2014 using the average of the nearest neighbours. Mathematically, the autocorrelation function provides a measure of the signal's characteristics in

 © Pychtar Mykola, Lytvyn Myroslava, 2026

the time domain, which is used to describe the relationship between the signal's value at one point in time and its value at another point in time (Fig. 1).

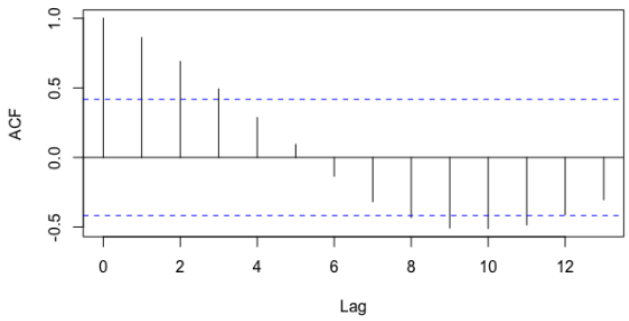

Fig. 1. Autocorrelation

To analyse trends in morbidity, time series analysis was used, with time series classified as stationary or non-stationary. Non-stationary data, which is typical of medical indicators, is presented as the sum of the trend $f(t)$ and the stationary process

$$X_t : Y_t = f(t) + X_t, t \in \{0, 1, ..., n-1\}.$$

The ARIMA (Autoregressive Integrated Moving Average) model has been selected as the primary forecasting method; it is characterised by the following parameters (p, d, q). Mathematically, the model for a non-stationary time series takes the form():

$$\Delta^d X_t = c + \sum_{i=1}^{p} a_i \Delta^d X_{t-1} + \sum_{j=1}^{q} b_j \epsilon_{t-j} + \epsilon_t,$$

where Δ_d is the difference operator of order d, and ϵ_t is a "white noise".

The study is based on morbidity data for the city of Kyiv for the period 1995–2016. The key indicators identified are:

- p is the percentage of children among all patients with diabetes mellitus (DM).
- fp is the proportion of children among those newly diagnosed with diabetes. To analyze the dataset, R programming language for statistical computing and data visualization was used.

Statistical analysis revealed significant variability in the fp value (coefficient of variation 36.0%) compared with p (19.6%).

Modelling and forecasting results. Two forecasting models have been developed for the p indicator. The first model is $ARIMA(0, 2, 1)$ model: The coefficients were selected using the Akaike Information Criterion (AIC). The model indicates an upward trend in the proportion of sick children over the next five years.

The second model was a linear regression model; the least-squares method (LSM) was used to estimate the trend parameters. The analysis of the parameters revealed that the coefficients were statistically significant ($p - value < 0.05$); however, the model explains only 20% of the variation in the series ($R = 0.2091$)

A comparison of the results reveals a discrepancy: the ARIMA model predicts a further rise in incidence, whilst linear regression indicates a slight downward trend, highlighting the advantage of dynamic models for short-term forecasting of health indicators.

2. Diabet prediction based on medical records

This section is devoted to the analysis of the diabetes prevalence dataset in women with 9 other variables: number of pregnancies, blood glucose, blood pressure, body mass index (BMI), genetic predisposition to diabetes, insulin, age, and presence of diabetes from (Smith et al., 1988). The dataset was collected by the National Institute of Diabetes and Digestive and Kidney Diseases. The purpose of the dataset is to diagnose whether a patient has diabetes or not based on certain diagnostic measurements included in the dataset. All patients in this dataset are women aged at least 21 years.

The contrast matrix (last column) in Fig. 2 visually demonstrates differences in diabetes incidence across several characteristics, namely the number of pregnancies, glucose level, body mass index (BMI), and age. However, these observed differences require further statistical verification to confirm their significance.

To analyze the relationships between variables in more detail, a matrix of pairwise correlation coefficients is constructed. This matrix reflects the degree of linear dependence between each pair of factors. The lower triangular part of the matrix contains the numerical values of the correlation coefficients, while the upper part provides a visual representation using circles of varying intensity: with blue shading indicating a positive outcome (with diabetes) and red shading indicating a negative outcome (without diabetes).

The Student's t-test is one of the basic statistical methods used to compare means between groups of data and to identify statistical differences. Comparing means allows you to determine whether the selected groups are significantly different in terms of the indicator or outcome under study. In the context of medical research, comparing mean values between groups, such as patients and healthy individuals, can help identify risk factors that contribute to the development of diseases. This approach allows you to determine which indicators or properties of the body differ between groups and may be associated with the onset and progression of specific diseases. This information can be important for developing

strategies to prevent and manage disease risks. Suppose we want to compare two sample means: X_1, calculated from a sample with standard sd_1 deviation and sample size n_1 and X_2 calculated from a sample with standard deviation sd_2 and sample size n_2. Then, we formulate the following hypotheses:

a) H_0: In the population, there is no difference between the mean values.

b) H_1 The means in the population are not equal (the alternative hypothesis).

Fig. 2. Matrix of contrasts and pairwise coefficients of correlations depending on whether there is diabetes

First, we assume that the null hypothesis H_0 holds. This means that in the population, there is indeed no difference between the mean values X_1 and X_2 . If this is the case, then in repeated experiments (drawing samples and calculating the difference between the two sample means), the difference $X_1 - X_2$ should be close to zero. This is compared with the tabulated value of the Student's t-distribution with n_1+n_2-2 degrees of freedom, which depends on the chosen significance level. In R, we will use the function t.test and compare the obtained p-value with the significance level 0.05 (Tab. 1). If the p-value is greater than 0.05, there are no grounds to reject the null hypothesis at the 0.95 confidence level (the means are statistically equal). Otherwise, the null hypothesis is rejected (the means differ with probability 0.95).

Table 1

Statistical significance of the difference between healthy and sick people

Factor	t	p**-value**	**Is the difference stat.significant?**
Pregnencies	-5.907	3.411e-09	yes
Glucose 1	-13.752	2.2e-16	yes
BloodPreassure	-1.7131	0.08735	no
Skin Thickness	-1.9706	0.05	no
Insulin	-3.3009	0.001047	yes
BMI	-8.6199	2.2e-06	yes
DPF	-4.5768	6.1e-06	yes
Age	-6.9207	1.202e-11	yes

For the classification of diabetes, three machine learning methods were applied: k-Nearest Neighbors, Random Forest, and Logistic Regression.

The k-Nearest Neighbors (k-NN) method is one of the simplest machine learning algorithms used for classification and regression. It is based on the assumption that similar objects have similar outcomes. In this method, there is no explicit training process. The algorithm simply stores all the training data along with their labels (in the case of classification) or values (for regression). To predict the class or value of a new object, the algorithm computes the distances between this object and all objects in the training set. For applying the k-NN method, all values were normalized (for each variable, the mean was subtracted and divided by the standard deviation). Next, the entire dataset was split into training and testing sets in an 80 – 20 % ratio. For the nearest neighbors method, we use the parameter k = 5. The knn() function is used to build the classification model, and a confusion matrix is constructed on the test set.

Random Forest is a commonly used machine learning algorithm that combines the outputs of multiple decision trees to achieve a single result. A random forest is an ensemble of decision trees, where each tree depends on a certain set of random variables. To apply the Random Forest method, all data are split into training and testing sets in an 80 – 20 % ratio. Using the randomForest() function, a classification model is created, and a confusion matrix is constructed.

Logistic regression is a statistical method used to model the relationship between a dependent variable (response) and one or more independent variables (predictors). Its primary purpose is to predict the probability of an event with a binary outcome, for example, success or failure, a positive or negative diagnosis, and so forth. The dependent variable (Y)

in logistic regression is categorical, most commonly binary ($Y \in \{0, 1\}$). In our case, 0 indicates no Diabetes, 1 indicates the presence of Diabetes.

The model estimates the probability p that the event will occur: $p = P(Y = 1|X)$.

To transform probabilities into a form suitable for linear modeling, logistic regression applies the logit function (the natural logarithm of the odds):

$$logit(p) = ln(p/(1-p)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k,$$

where $\beta_0, \beta_1, \ldots \beta_k$ are model parameters estimated from the data.

Using the glm() function, we obtained the following logistic regression model for our dataset:

$$logit(p) = -8.404 + 0.123X_1 + 0.025X_2 - 1.33X_3 + 0.001X_4 + 0.089X_6 + 0.945X_7 + 0.014X_8.$$

Based on this model, we classified individuals as either healthy or diabetic. The classification was performed using the following R command, which assigns an observation to the diabetic group if the predicted probability exceeds 0.5.

To evaluate the model's performance, a confusion matrix was generated using a table. This matrix provides a summary of correctly and incorrectly classified cases, allowing for an assessment of the model's accuracy.

3. Classification performance of quality metrics

Accuracy is one of the most widely applied quality metrics, defined as the proportion of correctly classified instances.

Sensitivity is used to assess the ability of a test or model to correctly identify positive cases. In other words, sensitivity indicates the proportion of truly positive cases that were correctly classified as positive. From the table, it can be seen that the method providing the highest sensitivity is Random Forest.

Specificity is used to evaluate the quality of a test or model. While sensitivity measures how well the test detects positive cases, specificity reflects how well it identifies negative cases. In other words, it is the ability of the test to correctly recognize the absence of a certain feature or disease. According to the table, Logistic Regression demonstrates the highest specificity.

Precision characterizes the degree of consistency of results when measurements or experiments are repeated. In other words, precision shows how closely the results of repeated measurements of the same quantity agree with each other. The highest level of precision is achieved by Random Forest.

Negative Predictive Value (NPV) indicates how reliable a negative test result is. In other words, NPV answers the question: "If the test result is negative, how likely is it that the individual truly does not have the condition?" The highest NPV is achieved by Logistic Regression (see Tab. 2).

Table 2

Statistical significance of the difference between healthy and sick people

	k-NN	Random Forest	Logistic Reg
Accuracy	0.7208	0.7662	0.78255
Sensitivity	0.8617	0.8842	0.58209
Specificity	0.5000	0.5763	0.89
Precision	0.7297	0.7706	0.73933
NPV	0.6977	0.7556	0.79892

Discussion and conclusions

This paper examined and analyzed type 2 diabetes mellitus. First, the prevalence of diabetes in Kyiv from 1995 to 2015 was considered. Using ARIMA processes and a regression model, the short-term incidence of diabetes in Kyiv was forecasted. Then we investigated and described another dataset consisting of one categorical variable (presence of diabetes) and eight numerical variables. To explore the data in more detail, graphical visualization methods were used. This allowed us to assess the distribution of each variable and detect potential outliers. To identify relationships between variables and their dependence on the presence of diabetes, a correlation matrix was built. Then, to test whether the mean values of indicators differ between diabetic and healthy individuals, a statistical test – Student's t-test – was applied. For patient classification into groups, three machine learning methods were used: k-nearest neighbors, logistic regression, and Random Forest. Logistic regression achieved the highest classification accuracy (0.78). The method with the highest sensitivity was Random Forest (0.88). Logistic regression achieved the highest specificity (0.89).

The conducted analysis shows that diabetes incidence is influenced by multiple factors, with glucose level, BMI, age, and number of pregnancies being statistically significant predictors. The time series modeling for Kyiv revealed that the ARIMA model is more suitable for short-term forecasting of non-stationary medical data, while linear regression provides less reliable results.

The comparison of classification methods demonstrated that Random Forest achieves the highest sensitivity and precision, making it more effective for detecting diabetes cases. Logistic regression showed higher specificity and negative predictive value, indicating better performance in identifying healthy individuals. The k-NN method yielded comparatively lower accuracy.

Overall, the results confirm that combining statistical methods with machine learning improves both forecasting and early diagnosis of diabetes, which is important for preventive healthcare and decision-making.

Authors' contribution: Mykola Pychtar – conceptualization, methodology; Myroslava Lytvyn – software, empirical data collection and validation, empirical research, analysis of sources, preparation of literature review.

Acknowledgments. The authors are grateful to the Armed Forces of Ukraine. Without their heroic contributions, this and many other studies would not be possible.

Sources of funding. This study did not receive any grant from a funding institution in the public, commercial, or non-commercial sectors.

References

Aguilera-Venegas, G., López-Molina, A., Rojo-Martínez, G., & Galán-García, J. L. (2023). Comparing and tuning machine learning algorithms to predict type 2 diabetes mellitus. *Journal of Computational and Applied Mathematics*, 427, 115115. https://doi.org/10.1016/j.cam.2023.115115

IDF. (2024). *The the diabetes atlas.* https://diabetesatlas.com.ua/ua/kyiv.

Iftikhar, K., Javaid, N., Ahmed, I., & Alrajeh, N. (2025). A novel explainable deep learning framework for accurate diabetes mellitus prediction. *Applied Sciences*, 15(16), 9162. https://doi.org/10.3390/app15169162

Kangra, K., & Singh, J. (2023). Comparative analysis of predictive machine learning algorithms for diabetes mellitus. *Bulletin of Electrical Engineering and Informatics*, 12(3), 1728–1737. https://doi.org/10.11591/eei.v12i3.4412

Livinska, H., Skrypnyk, D., & Galán-García, J. L. (2024). Comparative study of machine learning algorithms for diabetes detection using binary data. *Bulletin of Taras Shevchenko National University of Kyiv. Physics and Mathematics*, 78(1), 119–127. https://doi.org/10.17721/1812-5409.2024/1.23

Olisah, C. C., Smith, L., & Smith, M. (2022). Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective. *Computer Methods and Programs in Biomedicine*, 220, 106773. https://doi.org/10.1016/j.cmpb.2022.106773

Orlova, N., Tkachenko, O., Herasymiuk, K., Palamar, I., & Sereda, N. (2026). Epidemiological situation of type 2 diabetes mellitus in ukraine: challenges and priorities for public health management. *Ukraine. Nation's Health*, 1, 28–37. https://doi.org/10.32782/2077-6594/2026.1/03

Smith, J., Everhart, J., Dickson, W., Knowler, W., & Johannes, R. (1988). Using the adap learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the symposium on computer applications and medical care* (pp. 261–265). IEEE Computer Society Press. https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

Отримано редакцією журналу / Received: 23.10.25
Прорецензовано / Revised: 13.01.26
Схвалено до друку / Accepted: 16.04.26
Опубліковано / Published: 05.06.26

Микола ПИХТАР, канд. пед. наук, доц.
ORCID ID: 0000-0003-4567-8901
e-mail: nppihtar@gmail.com
Український державний університет імені Миxайла Драгоманова, Київ, Україна

Мирослава ЛИТВИН, студ.
ORCID ID: 0009-0006-6792-6053
e-mail: myroslava.lytvyn.25@kse.org.ua
Київська школа економіки, Київ, Україна

СТАТИСТИЧНИЙ АНАЛІЗ ФАКТОРІВ, ЩО ВПЛИВАЮТЬ НА ЗАХВОРЮВАНІСТЬ ЦУКРОВИМ ДІАБЕТОМ

Серйозність проблеми діабету фахівці пояснюють рівнем поширення хвороби. У 2021 р. приблизно 537 млн людей на всій Землі хворіли на цукровий діабет, що становить приблизно 6 % населення. З огляду на це дослідження та аналіз даних за динамікою захворювання цукрового діабету є актуальними як з позиції побудови прогнозу захворюваності, так і з позиції можливості державного регулювання та запобігання поширенню хвороби. Метою проєкту є статистичний аналіз захворюваності на цукровий діабет, побудова моделей динаміки захворюваності для міста Києва, а також аналіз впливу таких показників у жінок: кількість вагітностей, рівень глюкози в крові, кров'яний тиск, індекс маси тіла (ІМТ) тощо, на цукровий діабет. До цих змінних застосовано методи класифікації для побудови прогнозних значень за допомогою мови програмування R. Відповідно до поставленої мети в роботі розв'язано такі основні завдання: відшукання у відкритому доступі даних щодо цукрового діабету; опрацювання відомих підходів до аналізу таких даних; обчислення базових основних статистичних показників захворюваності та візуалізація даних за допомогою мови програмування R; знаходження тренду захворюваності; побудова моделі та прогнозних значень на основі процесів ARIMA та регресійного аналізу. За допомогою критерію Стьюдента здійснено перевірку на відмінності в середньому таких показників для здорових і хворих на цукровий діабет, як кількість вагітностей, рівень глюкози в крові, кров'яний тиск, ІМТ тощо. Прогнозування захворювання на діабет реалізовано за допомогою трьох алгоритмів машинного навчання, а саме: k-NN; Random forest; логістичної регресії. Виконано обчислення мір якості алгоритмів і порівняння точностей класифікації.

Ключові слова: ***часовий ряд, цукровий діабет, прогнозування та моделювання, k-NN, Random forest, логістична регресія, точність алгоритмів.***

Автори заявляють про відсутність конфлікту інтересів. Спонсори не брали участі в розробленні дослідження; у зборі, аналізі чи інтерпретації даних; у написанні рукопису; в рішенні про публікацію результатів.

The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; in the decision to publish the results.