Statistical Analysis of Diabetes Risk Factors Using 2*2Contingency Tables in R

1- Introduction:

Diabetes is a chronic disease with a growing prevalence worldwide. Understanding the behavioral and physiological indicators associated with diabetes is essential for early detection and prevention. This report aims to examine the association between several categorical risk factors and diabetes diagnosis using two-way contingency tables and various statistical measures such as:

- Proportion (π) in each group
- Difference in proportions
- Relative Risk (RR)
- Odds Ratio (OR)
- Chi-Square Test of Independence
- Residual and Standardized Residual analysis

The analysis was performed using R, and the dataset was obtained from an open-source medical database available on Kaggle.

2- Dataset Description

The dataset includes 521 individual cases, each representing a patient record with various symptoms and diagnostic information. The variables are mostly categorical (Yes/No), and the primary outcome variable is `Class`, which indicates whether the person has diabetes (Positive) or not (Negative).

The dataset contains 17 variables, of which 15 are categorical. In this analysis, we focus on the following five key variables:

- **Polyuria** (Frequent urination)
- **Polydipsia** (Excessive thirst)
- Sudden Weight Loss
- Weakness
- Visual Blurring

3-Methodology

To analyze the association between each categorical risk factor and diabetes diagnosis, we used the following statistical procedures:

1. Two-Way Contingency Tables

For each variable, a 2x2 table was created comparing the presence/absence of the risk factor with the diabetes diagnosis (Positive / Negative).

2. Chi-Square Test of Independence

This test was used to assess whether the two variables are statistically independent. A p-value of less than 0.05 was considered statistically significant.

3. **Proportions and Differences**

We computed the proportion of diabetes cases among individuals with and without each risk factor (π_1 and π_2), and the difference between them ($\pi_1 - \pi_2$).

4. Relative Risk (RR)

RR was calculated to assess how much more likely diabetes is among those with the risk factor.

5. Odds Ratio (OR)

OR was used to estimate the odds of diabetes in one group compared to the other.

6. Residual Analysis

Raw and standardized residuals were analyzed to detect which cells contributed most to any significant association.

7-

7. Data Visualization

For each relationship, a relative bar plot (using ggplot2 in R) was generated to visualize the distribution of diabetes diagnosis across levels of the predictor.

9-****Predictive Modeling using Logistic Regression***

9- ## 📊 Integration of Predictions into the Full Dataset

After developing and evaluating the logistic regression model, we applied it to the entire dataset (521 records). Two new columns were created and appended to the main data frame:

- `Prediction`: Contains the model's binary classification of each patient as either "Positive" or "Negative" for diabetes.

- `Predicted Probability`: Represents the estimated probability (between 0 and 1) that the patient is diabetic, based on the logistic regression model.

These columns enhance the interpretability of the model, allowing for further filtering, visualization, and in-depth analysis. This step also allows us to export the full enriched dataset for practical or academic use.

10- Model Performance and Confusion Matrix

11-## 📊 Visualization of Model Discrimination: ROC Curve

All calculations and visualizations were performed using R programming language.