"Ss. Cyril and Methodius University" – Skopje
Faculty of Computer Science and Engineering

Analysis of nutrition and general health status in multiple countries of the World

---

Predictive modeling of diabetes prevalence based on nutrition and health indicators across multiple countries

Hristina Sekuloska
211236

Anastasija Janakjievska
213120

Skopje, 2025

Introduction

Diabetes is a growing global health issue influenced by genetic, demographic, and lifestyle factors. This project aims to analyze international data to identify key factors affecting diabetes prevalence ages 20–79 and to create a dataset suitable for machine learning and risk prediction by country.

Data Sources

We used data from three main sources:

- IDF Diabetes Atlas – provided diabetes prevalence rates (ages 20–79) for each country, which were used as the target variable. *(https://diabetesatlas.org/)*
- World Bank – offered indicators like population, obesity rates, and food availability. *(https://data.worldbank.org/)*
- FAO – contributed nutritional data such as sugar, fat, and calorie intake. *(https://www.fao.org/faostat/en/#data)*

As data came from different years, we selected the most recent available value for each indicator.

Data Cleaning and Preprocessing

After unifying country names, duplicates were removed by keeping the most complete record. Countries with too many missing values were excluded to ensure better data quality. Remaining missing values were filled using median imputation. Although the model hasn't been implemented at this stage, due to the wide range of variable scales (e.g. population vs. vitamin intake), future modeling will require normalization or standardization (e.g. using StandardScaler from scikit-learn).

Feature Reduction and Indicator Selection

At the initial stage, the dataset included over 30 indicators covering various domains such as demographics, nutrition, economic conditions, food access, and health statistics. However, including all of these features in a predictive model could lead to several challenges — increased model complexity, higher risk of overfitting, and multicollinearity among features, which may affect model stability.

To address this, a feature selection process was applied using the following steps:

- Correlation analysis was performed to assess the relationship between each indicator and the target variable.
- Indicators with a high percentage of missing values were removed from consideration.
- A heatmap visualization was used to identify and eliminate features that were strongly correlated with each other.
- Finally, theoretical selection was applied by keeping only those indicators that are supported in scientific literature as relevant to diabetes prevalence.

As a result, the dataset was reduced to 9 carefully selected independent indicators, along with one target column representing the percentage of Diabetes in the population aged 20 to 79.

Feature Selection

Based on initial analysis and domain knowledge, the following 9 indicators were selected as explanatory variables for the model:

1. Population, total – Total population of the country
2. Obese adults (18+), million – Number of obese adults
3. Vegetables, other – Consumption of vegetables (excluding potatoes)
4. Prevalence of overweight (%) – Percentage of overweight adults
5. Per capita food supply variability – Variability in food supply per person
6. Energy supply (kcal/cap/day) – Daily caloric intake per person
7. Sugar (Raw Equivalent) – Sugar consumption
8. Fat supply – Fat intake
9. Vitamin B6 supply – Availability of vitamin B6

These indicators were chosen based on theoretical relevance and correlation analysis, ensuring they provide diverse yet meaningful input for modeling diabetes prevalence.

Model Training

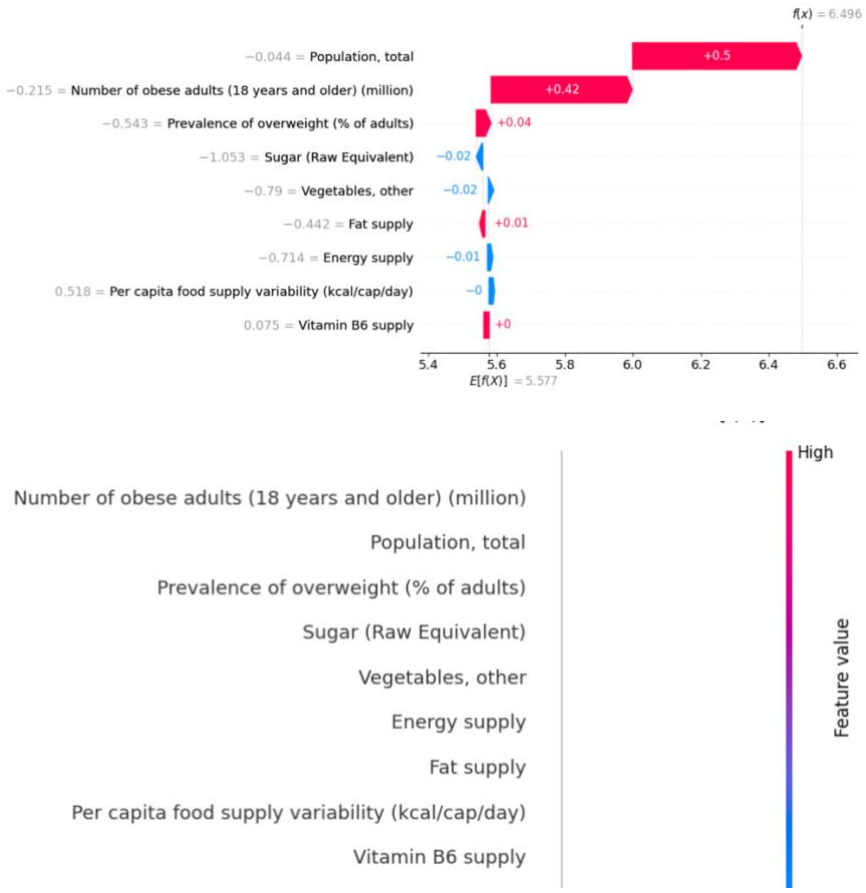
To train the model, we used a Random Forest Regressor applied on a selected subset of the most relevant features. Before training, we standardized the input data using StandardScaler to ensure consistent scaling across variables. Outliers were detected and removed using Isolation Forest, which helped improve the model's reliability.

We applied Leave-One-Out Cross-Validation (LOOCV) to evaluate performance, where each country was left out once as a test case. To stabilize the variance and improve predictions, we log-transformed the target variable. The model's performance was assessed using RMSE and R^2, both in the log-transformed and original scale.

This approach allowed us to build a stable and interpretable model, suitable for understanding patterns in diabetes prevalence across countries.

Results Analysis

The Random Forest model performed well on the log-transformed target, achieving and R^2 of 0.8752 and RMSE of 0.63, indicating strong predictive power. On the original scale, performance dropped slightly $R^2 = 0.6463$, which is expected due to skewed data. Most countries had predictions close to actual values, though deviations appeared in cases with extreme diabetes rates. These outliers suggest potential local factors or data limitations. Overall, the model effectively captures global trends and highlights key predictors of diabetes prevalence.



Conclusion

The analysis uncovered several important findings. Firstly, the availability of calories, sugar, and fat is positively linked to the prevalence of diabetes, reinforcing their role as risk factors. Obesity and overweight rates also emerged as strong predictors, consistent with medical studies connecting BMI to type 2 diabetes. Conversely, certain nutritional factors such as vitamin B6 intake and vegetable consumption appeared to have a protective effect, indicated by negative coefficients in the model. The model employed is both interpretable and stable, demonstrating reliable predictive performance. It provides a solid foundation for public health analysis and risk assessment across different countries. Future directions include expanding the dataset over time, exploring more advanced models to enhance accuracy, and utilizing clustering methods to categorize countries by their risk levels.