“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 colorectal cancer incidence based on nutritional and health indicators across multiple countries

Anastasija Janakjievska 213120

Skopje,2025

Project Overview

Colorectal cancer (CRC) is one of the leading causes of cancer mortality worldwide. This project analyzes nutritional and lifestyle factors influencing country-level CRC prevalence using real-world data and modern machine learning techniques. By comparing models and clustering results, key risk factors are identified and countries are categorized by health profile.

Data Sources

Data were collected from four reputable global repositories:

- World Bank Open Data (demographic & health indicators)
- FAOSTAT (nutritional data: protein, sugar, fruit, honey)
- WHO Global Health Observatory (physical activity & tobacco use)
- IARC Global Cancer Observatory (official CRC incidence data)

Data Preprocessing

- Merged tables on the “Country” field after harmonizing country names
- Excluded countries with >30% missing data
- Imputed remaining missing values using median substitution
- Removed **eight** outlier countries via an Isolation Forest for greater model stability

Modeling & Comparison

Hyperparameters for both models were optimized using **RandomizedSearchCV**

XGBoost Regressor

- Best parameters: subsample=0.7, n_estimators=300, min_child_weight=3, max_depth=5, learning_rate=0.1, gamma=0, colsample_bytree=1.0
- LOOCV performance:
  - RMSE (log-target): 0.4631
  - R^2 (log-target): 0.8309
  - RMSE (original scale): 15.1972
  - R^2 (original scale): 0.7537

Random Forest Regressor

- Best parameters: n_estimators=300, min_samples_leaf=1, max_features=0.5, max_depth=15
- LOOCV performance:
  - RMSE (log-target): 0.4766
  - R^2 (log-target): 0.8209

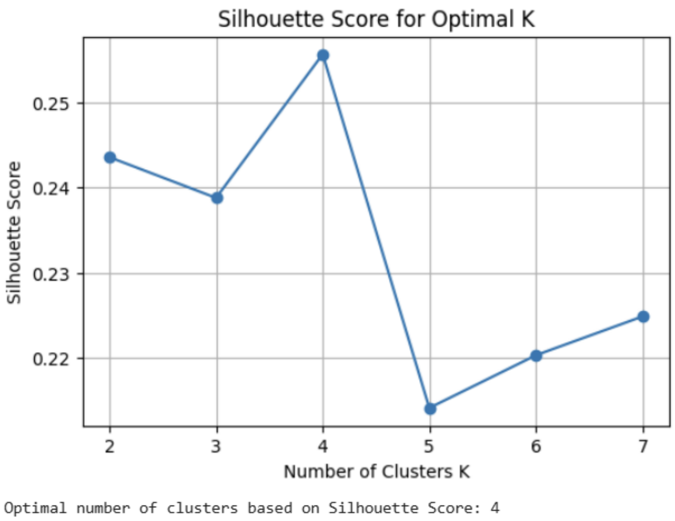
- RMSE (original scale): 16.6246
- R² (original scale): 0.7052

Across all metrics, the XGBoost model outperformed Random Forest, making it the preferred choice for predicting CRC incidence.

Cluster Analysis

KMeans clustering (k=4, based on silhouette score) was applied to eight indicators:

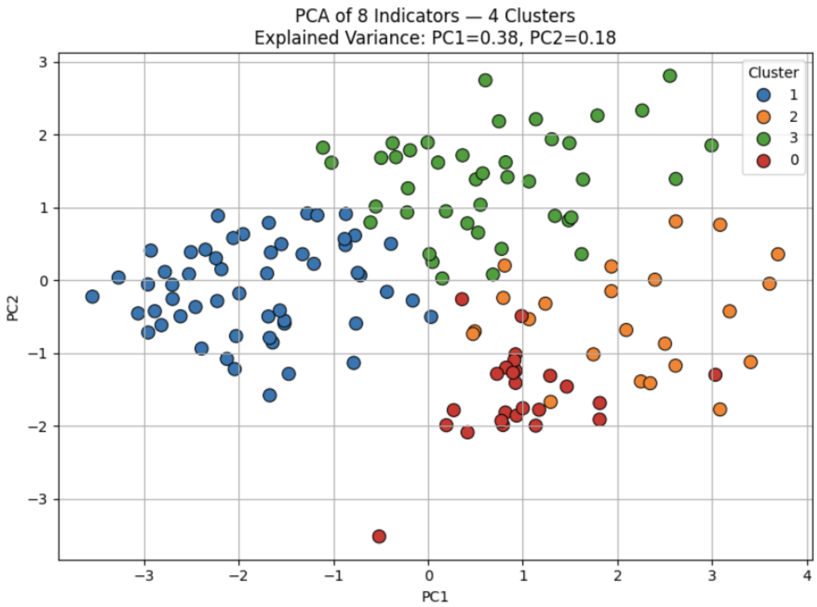
- Average supply of animal-origin protein (g/capita/day)
- Total alcohol consumption per capita (liters)
- Honey supply
- Oranges & mandarins supply
- Sugar (raw equivalent) supply
- Average protein supply (g/capita/day)
- Insufficient physical activity (%)
- Current tobacco use prevalence (%)



```
Optimal number of clusters based on Silhouette Score: 4
```

Cluster summaries:

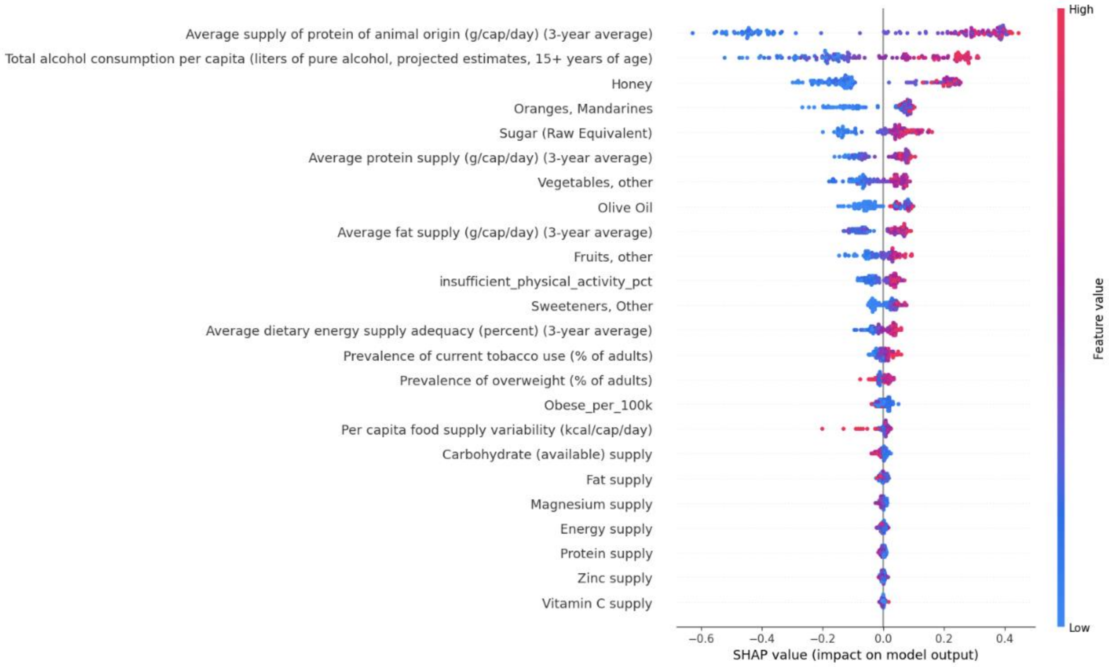
- **Cluster 0 (25 countries):** Moderate protein & fruit intake; high alcohol & tobacco use; moderate obesity.
- **Cluster 1 (55 countries):** Lowest values across all indicators – healthiest profile.
- **Cluster 2 (23 countries):** Very high protein & sugar intake; increased inactivity & obesity.
- **Cluster 3 (42 countries):** Extreme obesity & inactivity; highest sugar consumption.



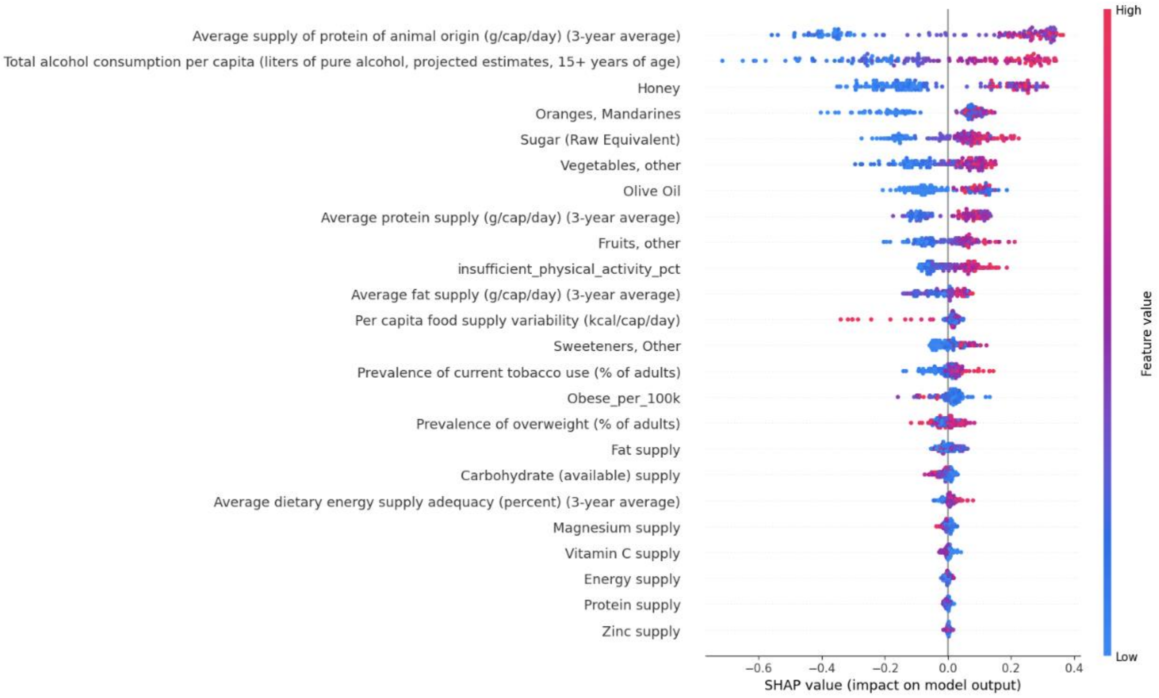
SHAP Analysis

To interpret feature contributions, SHAP (SHapley Additive exPlanations) values were computed for both models, revealing how each input factor increases or decreases predicted CRC incidence.

- ***Random Forest SHAP:***



- ***XGBoost SHAP:***



Conclusion

Nutritional and lifestyle factors—particularly sugar intake, physical inactivity, and obesity—significantly explain global CRC prevalence variations. The XGBoost model demonstrated superior predictive performance. Cluster analysis effectively stratified countries by health risk profile, and SHAP analysis provided transparent insights into the most influential factors. Future work may extend temporal analyses and explore additional advanced modeling approaches.