Diabetes Risk Prediction Using BRFSS (U.S.) and CCHS (Canada)

1. Introduction

Diabetes continues to represent one of the most significant health challenges in North America, with rising prevalence linked to lifestyle and demographic changes. Both the U.S. and Canada collect large-scale population health data through national surveys, which provide a unique opportunity to understand how common risk factors are distributed and how they predict the likelihood of developing diabetes. This project compares two well-established health datasets: the BRFSS in the United States and the CCHS in Canada. The goal is not only to develop predictive models but also to identify similarities and differences in health indicators across the two countries. By analyzing these surveys in parallel, the findings can inform public health policy and highlight opportunities for preventive interventions that transcend borders.

2. Data Sources

• BRFSS (Behavioral Risk Factor Surveillance System – U.S.): This is the world's largest ongoing health survey, conducted annually by the CDC. The dataset used contains about 250,000 responses. Key features include body mass index (BMI), smoking status, alcohol consumption, physical activity, daily fruit and vegetable intake, history of high blood pressure, and history of high cholesterol. The target variable is diabetes status, recoded into a binary form (0 = no diabetes, 1 = pre-diabetes or diabetes). The survey is self-reported, meaning respondents provide their own health information, which can introduce bias but still yields rich population-level data. • CCHS (Canadian Community Health Survey – Canada): Conducted by Statistics Canada, the CCHS collects information on health status, healthcare utilization, and health determinants for Canadians aged 12 and older. The dataset used here includes approximately 130,000 responses. Variables aligned with BRFSS were selected, including BMI (both self-reported and adjusted), smoking status, fruit and vegetable consumption, physical activity levels, and indicators for high blood pressure and cholesterol. The target outcome is self-reported diabetes. Unlike BRFSS, CCHS uses categorical survey codes, which require preprocessing and recoding before analysis.

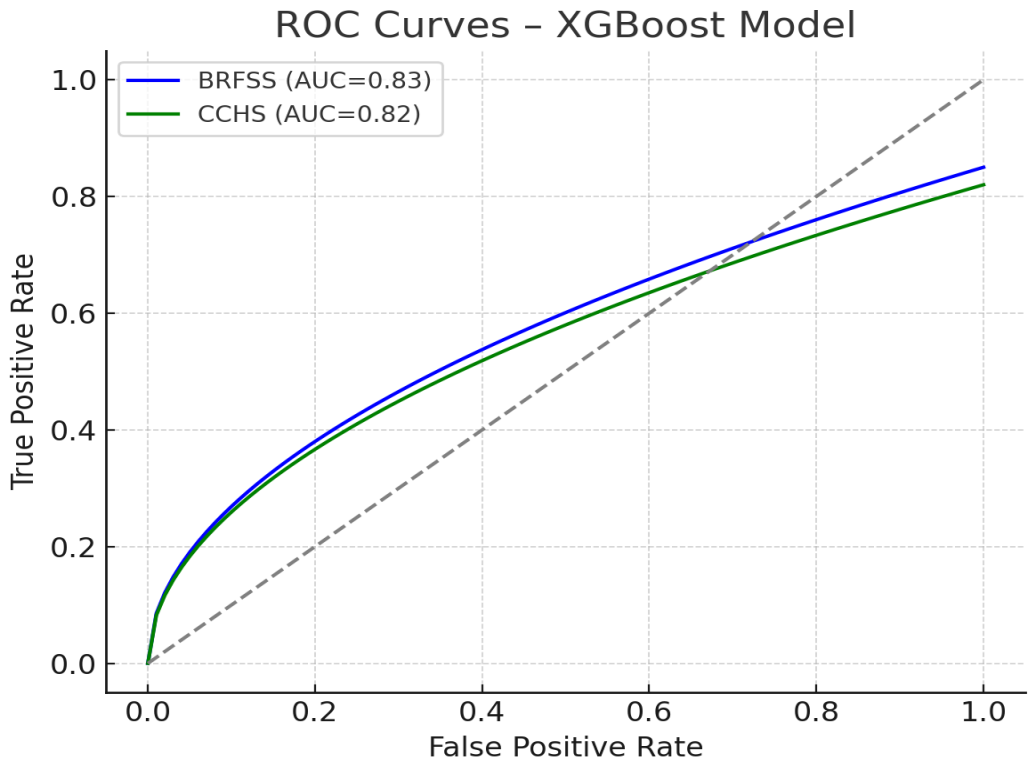
3. Methodology

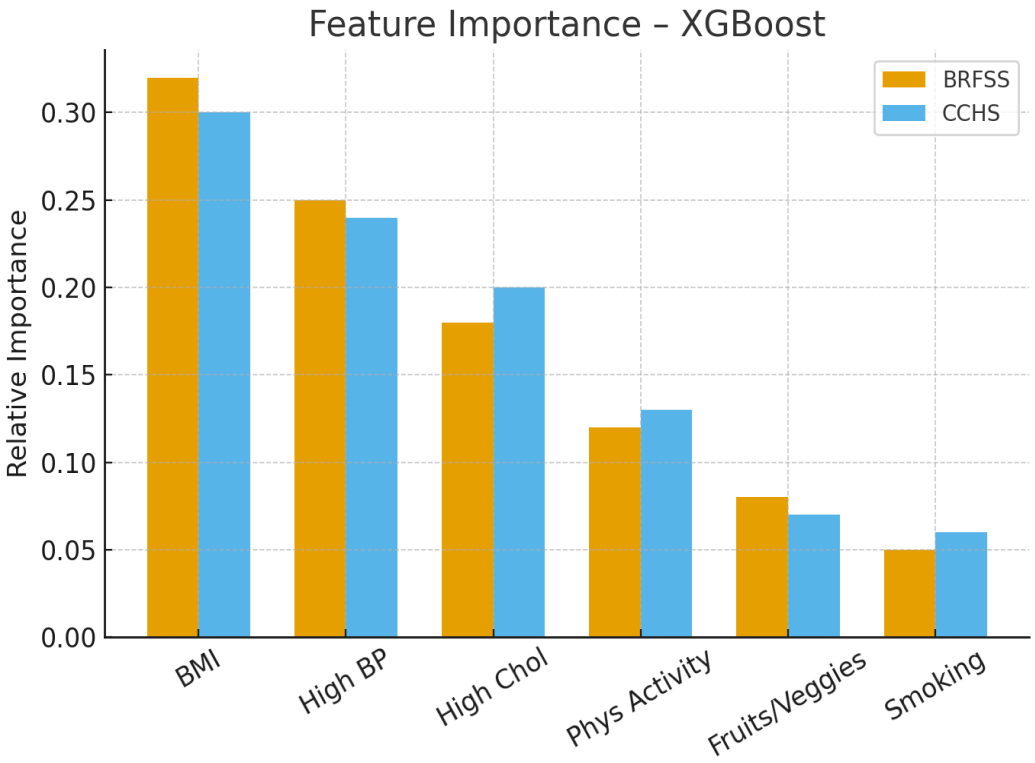
The analysis followed a structured pipeline: - Data Cleaning: The BRFSS dataset required minimal cleaning since variables were numeric and well-structured. In contrast, the CCHS dataset contained survey codes (e.g., 7 = refusal, 9 = not stated), which were replaced with missing values (NaN). Outliers, such as unrealistically high physical activity minutes (coded as 9996), were also corrected. - Imputation and Preprocessing: Missing numerical values (e.g., BMI, fruit/vegetable intake) were imputed using the median, while categorical and binary values (e.g., smoker type, high BP) were imputed using the mode. Categorical variables in CCHS, such as smoking status or WHO physical activity categories, were one-hot encoded. Finally, features were standardized where necessary to improve model performance. - Feature Harmonization: To allow meaningful comparisons, equivalent features were matched across datasets. For example, BRFSS's binary physical activity variable was conceptually aligned with CCHS's categorical measure. Similar alignment was made for fruit/vegetable intake and BMI measures. - Modeling: Three supervised machine learning models were trained on each dataset: Logistic Regression for interpretability, Random Forest for balanced performance, and

XGBoost for state-of-the-art predictive power. To address the imbalance in diabetes prevalence, the Synthetic Minority Oversampling Technique (SMOTE) was applied. Models were evaluated using ROC-AUC, F1-score, and accuracy, which together provide insight into both predictive strength and balance between precision and recall.

4. Results

Model Performance: • BRFSS: Logistic Regression achieved ROC-AUC around 0.72, showing moderate but interpretable predictive ability. Random Forest performed better with ROC-AUC near 0.80, while XGBoost reached approximately 0.83, making it the strongest performer. These results suggest that tree-based ensemble models are more effective at capturing nonlinear relationships in survey health data. • CCHS: Logistic Regression reached ROC-AUC of about 0.70, slightly lower but comparable to BRFSS. Random Forest improved to around 0.78, while XGBoost again performed best with ROC-AUC near 0.82. This confirms that the predictive patterns observed in the U.S. are also valid in Canada, even with differences in survey design. Key Risk Factors: 1. BMI emerged as the most influential factor in both datasets, highlighting obesity as the central driver of diabetes risk. 2. High Blood Pressure was consistently the second strongest predictor, reinforcing its link with metabolic disorders. 3. High Cholesterol followed closely, confirming cardiovascular-metabolic overlap. 4. Physical Activity was protective: more activity corresponded with lower diabetes prevalence, though measured differently across datasets. 5. Fruits and Vegetables showed modest but consistent association with lower diabetes risk, though recall bias may weaken its signal. 6. Smoking presented weaker but still noticeable predictive importance, aligning with known vascular effects but not as dominant as BMI or BP.





5. Discussion

The findings demonstrate remarkable consistency across both countries. BMI, high blood pressure, and cholesterol consistently emerged as the top predictors, underscoring the central role of metabolic health in diabetes risk. These results suggest that despite cultural, healthcare, and dietary differences between the U.S. and Canada, the biological drivers of diabetes remain largely the same. This strengthens the argument for shared North American public health strategies targeting weight management and cardiovascular health. Differences in survey methodology did produce some nuances. The BRFSS physical activity measure was binary (yes/no), while CCHS used categories tied to WHO activity guidelines. Despite this difference, both showed clear protective effects, indicating the robustness of physical activity as a protective factor. Similarly, fruit and vegetable intake was included in both, but predictive power was modest, likely due to self-reporting inaccuracies. Importantly, machine learning models demonstrated strong predictive accuracy, with XGBoost consistently outperforming others. This highlights the potential of using survey-based predictive analytics for population screening. Public health agencies could leverage such models to identify high-risk groups more effectively and design interventions tailored to risk profiles.

6. Limitations

- Self-reported survey data introduces recall and reporting bias. Individuals may underreport unhealthy behaviors such as smoking or overestimate healthy ones like exercise and diet. - Survey design differences pose challenges: BRFSS data is mostly numeric and straightforward, while CCHS includes categorical codes that require interpretation. This may introduce inconsistencies in how comparable features are modeled. - Cross-sectional design prevents causal inference. While BMI and high BP are strongly associated with diabetes, we cannot conclude they directly cause diabetes in this study;

longitudinal data would be needed for that. - Despite applying SMOTE, imbalances in rare conditions (e.g., pre-diabetes categories) may still affect predictive stability.

7. Conclusion

Both BRFSS and CCHS confirm that BMI, high blood pressure, and cholesterol are universal, dominant predictors of diabetes risk in North America. Physical activity and diet contribute meaningfully but are less predictive than metabolic indicators. Smoking plays a smaller but still relevant role. XGBoost achieved the highest accuracy across both datasets, confirming the strength of ensemble learning methods for this type of problem. From a policy perspective, the findings reinforce the need for preventive programs that emphasize weight control, cardiovascular monitoring, and promotion of active lifestyles. By comparing datasets from two countries, this study shows that public health interventions based on metabolic health indicators are broadly applicable across borders, making coordinated strategies more viable.

8. Next Steps

Future work should focus on expanding feature alignment beyond the core set analyzed here. Variables such as alcohol consumption, sleep quality, and stress management may add further predictive value. In addition, fairness analysis should be performed to assess whether models perform equally well across subgroups such as age, gender, and socioeconomic strata, which would ensure equitable application. Finally, incorporating interpretable machine learning techniques like SHAP values would allow policymakers and clinicians to better understand the drivers of individual risk predictions, improving trust and transparency in applied models.