A Comprehensive Machine Learning Framework for Predicting Health Risk Categories Using BRFSS Survey Data

Puspita Chowdhury
M.S. in Artificial Intelligence
Yeshiva University, New York, USA
Email: puspita.chowdhury@yu.edu

Abstract—**Chronic diseases represent a growing burden on global healthcare systems, making early identification of high-risk individuals a critical public health objective. Population-level survey datasets such as the Behavioral Risk Factor Surveillance System (BRFSS) contain valuable information on health behaviors and demographics but pose significant challenges due to missing values, noise, and class imbalance. This paper presents a comprehensive and leak-safe machine learning framework for predicting health risk categories using BRFSS survey data. The proposed framework includes rigorous data preprocessing, exploratory data analysis, feature selection, and systematic model evaluation. Multiple classification models are assessed, including a Dummy Classifier, Logistic Regression, and Random Forest. Experimental results demonstrate that Random Forest achieves the best overall performance, particularly in identifying high-risk individuals. The findings highlight the potential of machine learning-driven approaches to support scalable and data-driven public health decision-making.**

Index Terms—**Health Risk Prediction, BRFSS, Machine Learning, Public Health Analytics, Random Forest**

I. Introduction

The World Health Organization (WHO) has stated that cardiovascular disease, diabetes and obesity leading to morbidity and premature mortality globally; they are developed gradually over a long period and are shaped by an individual's lifestyle behaviors and the environment in which he or she lives (demographics/socioeconomic).

The CDC maintains the Behavioral Risk Factor Surveillance System (BRFSS) which, based on self-reported responses, is considered the largest and most comprehensive survey concerning health across all populations worldwide [5]. It collects self-reported data on health-related behaviors, preventive practices and chronic conditions from over a million people each year.

The BRFSS provides important information to public health officials; however, due to numerous methodological issues associated with BRFSS data, traditional statistical methods may not be effective in analyzing and interpreting the BRFSS data. Missing data, self-reported bias, high-dimensionality, and class imbalance are all problems that can arise with BRFSS data[3].

Machine Learning offers a viable alternative to the traditional statistical approach by allowing researchers to discover complex relationships between various features within the BRFSS data without having to make strong assumptions regarding the underlying data distribution[1]. The proposed research study will develop a comprehensive end-to-end Machine Learning (ML) system to mitigate analytical challenges posed by BRFSS data and ensure the overall rigor and reproducibility of the ML-based approach.

II. Problem Statement and Objectives

Many individuals with early-stage health risks remain undetected due to limitations in screening tools and reliance on symptom-based diagnosis. Traditional rule-based approaches lack adaptability and fail to scale to large, noisy population datasets.

The objectives of this study are:

- To preprocess and clean BRFSS survey data for machine learning analysis.
- To explore underlying patterns using exploratory data analysis.
- To build and compare predictive classification models.
- To select an optimal model based on validation performance.
- To evaluate generalization performance on unseen test data.

III. Related Work

In the field of health care, machine learning techniques such as logistic regression for predicting disease susceptibility, risk stratification, etc., have been applied successfully many times. The most common method is Logistic Regression because it is considered very easy to understand by both the doctor and patient and it has a good example of statistical theory as applied to epidemiological study designs[1].

However, logistic regression has limitations in its ability to account for nonlinearities and more complex relationships in the data due to its linear decision boundary [5]. Tree based ensembles, such as Random Forest, have historically performed very well when dealing with noisy data and when the target variable is continuous and/or not normally distributed.

Moreover, recent articles and research papers identify important considerations when working with survey based health data, in particular regarding self-reported health information that may provide clues about an outcome even before a person gives that response.

IV. Dataset Description

Data are from the BRFSS 2015 Data file (Best Practices Edition), which was obtained through Kaggle, where the CDC provides it [7]. Data contain representative responses from the entire US population and include respondent demographic variables, lifestyle behaviour and health, related indicators.

A. Target Variable

In order to classify respondents as either high risk or low risk to their health, respondents were assigned a ’Health Risk’ (HR) label (for both HR and LR) based on their responses. Any variables that coded directly for health outcomes were removed to avoid introducing ’target leakage’.

B. Feature Set

The final set of features used in the analysis will have demographic characteristics (age, sex, race), behavioural characteristics (e.g., smoking) and other quantifiable measures. All of the features selected were numeric, making them appropriate for use with machine learning models.

V. Data Preprocessing

Data preprocessing involved multiple steps to ensure reliability. Columns with more than 50% missing values were removed. Non-numeric attributes and unnecessary identifiers were excluded.

Self-reported health variables such as *genhlth*, *physhlth*, and *menthlth* were removed due to their direct relationship with the target label. Missing values were imputed using median imputation, and extreme outliers were removed. All preprocessing steps were implemented using machine learning pipelines to prevent information leakage.

VI. Exploratory Data Analysis

Exploratory Data Analysis (EDA) was conducted to examine feature distributions, class imbalance, and relationships among variables.

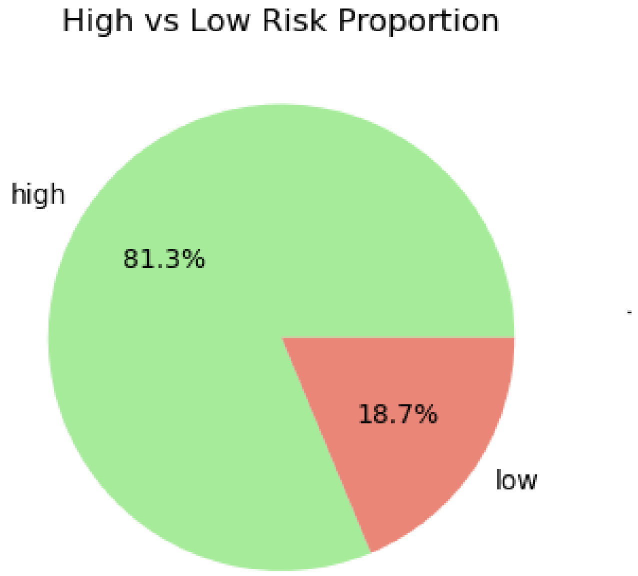


Fig. 1: Target Data

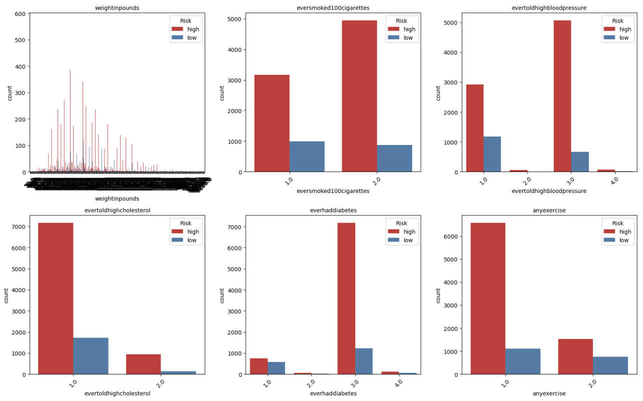


Fig. 2: Explanatory Data Analysis

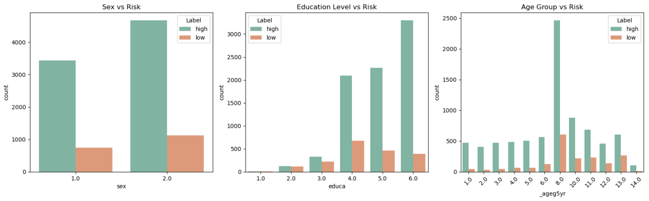


Fig. 3: Explanatory Data Analysis

The analysis revealed skewed distributions and significant class imbalance, motivating the use of evaluation metrics beyond accuracy.

VII. Feature Selection

Feature selection was guided by correlation analysis and domain knowledge. Highly correlated variables were examined to reduce redundancy and multicollinearity. Feature engineering was intentionally conservative to avoid data leakage and preserve interpretability.

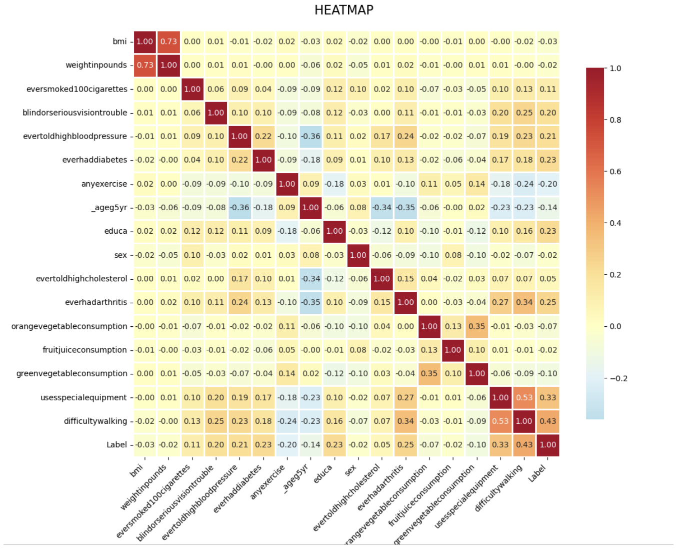


Fig. 4: Explanatory Data Analysis

VIII. Model Development

A stratified train-validation-test split was used to preserve class distributions. A Dummy Classifier served as a baseline to establish minimum expected performance.

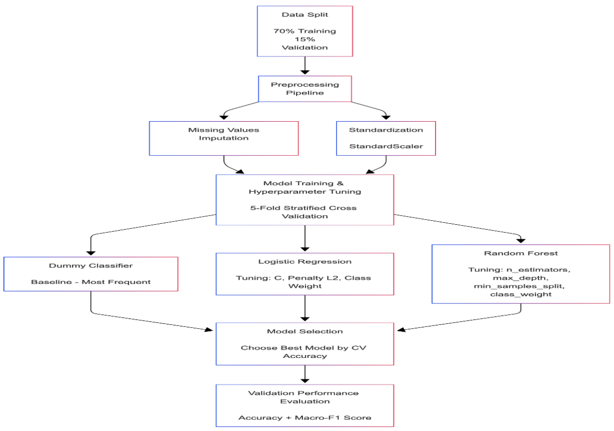


Fig. 5: Explanatory Data Analysis

A. Baseline Model: Dummy Classifier

Before training advanced machine learning models, a baseline classifier was established to provide a reference point for performance comparison. A Dummy Classifier was used as the baseline model, as it represents the simplest possible prediction strategy and helps determine whether more complex models provide meaningful improvement [4].

The Dummy Classifier does not learn from feature data. Instead, it generates predictions using predefined heuristics. In this study, the classifier employed the *most_frequent* strategy, which always predicts the majority class observed in the training dataset. Given the class imbalance present in the BRFSS data, this strategy reflects a realistic but naive baseline for population-level health risk prediction.

No hyperparameter tuning was performed for the Dummy Classifier, as it does not involve a learning process. The model was implemented using default settings provided by the scikit-learn library, with no trainable parameters. This ensures that the baseline performance is purely driven by class distribution rather than feature patterns.

The Dummy Classifier achieved a validation accuracy of 81.27%, reflecting the dominance of the majority class. However, the macro-averaged F1-score was only 44.83%, indicating poor performance in distinguishing between high-risk and low-risk individuals. These results highlight the limitations of naive prediction strategies and justify the use of more sophisticated machine learning models in subsequent stages of the framework.

B. Logistic Regression

Logistic Regression was employed as the first learnable classification model due to its strong theoretical foundation and widespread use in epidemiological and public health studies. Unlike the Dummy Classifier, Logistic Regression learns a probabilistic decision boundary based on input features, allowing it to model relationships between health behaviors and risk outcomes.

The model was implemented using L2 regularization to prevent overfitting and improve generalization. To address class imbalance in the BRFSS dataset, the *class_weight* parameter was set to *balanced*, ensuring that minority class errors were penalized more heavily during training. Hyperparameter tuning was conducted using cross-validation to identify the optimal regularization strength.

The best-performing Logistic Regression model achieved a validation accuracy of 80.41% and a macro-averaged F1-score of 72.69%. The optimal hyperparameters were a regularization strength of $C = 0.1$, L2 penalty, and balanced class weighting. Compared to the baseline Dummy Classifier, Logistic Regression significantly improved class discrimination, as reflected by the substantial increase in F1-macro score. However, its performance remained limited by the linear nature of the model, which restricts its ability to capture complex nonlinear interactions present in population-level health data.

C. Random Forest (Selected Model)

Random Forest was selected as the final classification model due to its superior performance and robustness when applied to large, noisy, and nonlinear datasets. As an ensemble learning method, Random Forest constructs multiple decision trees using bootstrap sampling and aggregates their predictions, thereby reducing variance and improving generalization [3].

To account for class imbalance, the *class_weight* parameter was set to *balanced_subsample*, allowing each tree to adapt to class distribution during training. Hyperparameter tuning was performed to optimize model complexity and prevent overfitting. The optimal configuration included 400 trees ($n\_estimators = 400$), a maximum tree depth of 10, and a minimum of 5 samples required to split an internal node.

The Random Forest model achieved the best overall validation performance, with an accuracy of 84.12% and a macro-averaged F1-score of 74.45%. Notably, the model attained a high-risk recall of 96%, indicating strong sensitivity in identifying individuals at elevated health risk. Although recall for the low-risk class was lower (31%), this trade-off is acceptable in preventive healthcare applications where minimizing false negatives for high-risk individuals is a priority.

Based on its strong predictive performance, robustness to noise, and ability to capture nonlinear feature interactions, Random Forest was selected as the final model for health risk classification.

IX. Validation Results and Model Selection

TABLE I: Validation Performance and Model Comparison

Model	Acc. (%)	F1 (%)	HR Rec. (%)	LR Rec. (%)
Dummy Classifier	81.27	44.83	–	–
Logistic Regression	80.41	72.69	–	–
Random Forest	**84.12**	**74.45**	**96**	31

TABLE II: Optimal Hyperparameters of Selected Model

Parameter	**Value**
Number of Trees ($n\_estimators$)	400
Maximum Depth (max_depth)	10
Minimum Samples Split	5
Class Weight	balanced_subsample

Based on validation performance, Random Forest was selected as the final model due to its superior accuracy, F1-score, and high-risk recall.

X. Test Results

The final Random Forest model was evaluated on a held-out test dataset.

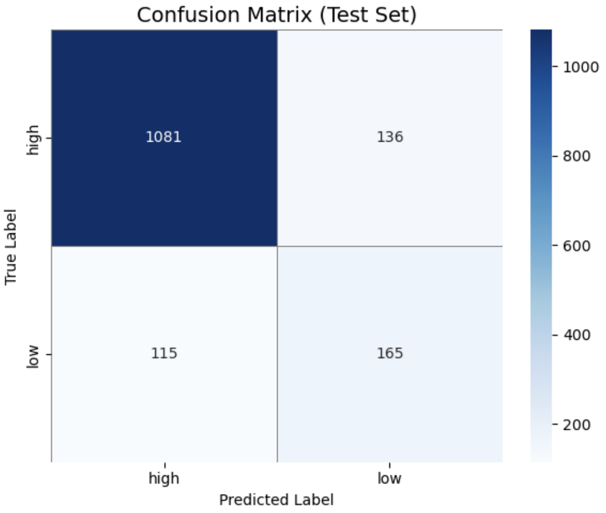
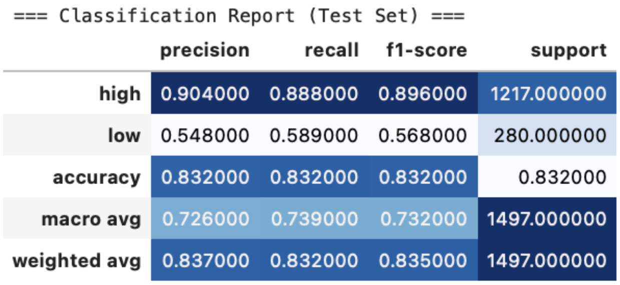


Fig. 6: Confusion matrix for Random Forest on test data



=== Classification Report (Test Set) ===

	precision	recall	f1-score	support
high	0.904000	0.888000	0.896000	1217.000000
low	0.548000	0.589000	0.568000	280.000000
accuracy	0.832000	0.832000	0.832000	0.832000
macro avg	0.726000	0.739000	0.732000	1497.000000
weighted avg	0.837000	0.832000	0.835000	1497.000000

Fig. 7: Classification Report

The model achieved a test accuracy of 83.83%, confirming strong generalization performance.

XI. Discussion

When it comes to predicting health risks at the population level, ensemble-based methods are more likely than linear ones to yield better predictions. The Random Forest algorithm efficiently models nonlinear relationships and complex interactions between demographic, behavioral, and lifestyle variables that would be difficult to model using linear systems[2]. Additionally, these types of interactions are particularly relevant to public health data sets such as BRFSS where multi-factorial relationships exist and where these risk factors tend to work together in some form of non-trivial additive fashion.

Random Forest models are also resilient to noise and missing values that are prevalent in most large health survey databases [5]. By combining the predictions of multiple models, Random Forest models reduce variance and dampen the effects of outliers and biases associated with self-reported behaviour. Thus, the enhanced robustness of Random Forest models results in increased stability and generalizability of predictions based upon previously unobserved input.

One additional and significant advantage of the Random Forest model is the high level of recall associated with the high-risk class, indicating a high sensitivity for detecting individuals with the greatest potential benefit from primary and secondary prevention services. In preventative healthcare systems this emphasis on identifying the highest risk individuals tends to have greater utility than maximizing overall accuracy, as a reduced number of missed high-risk cases produces a greater magnitude of delayed and increased health care cost burden. The emphasis on recall may, however, produce reduced recall for lower risk individuals.

XII. Ethical Considerations

When applying machine learning techniques to population-level health data ethical considerations need to be taken into account. While the BRFSS dataset used in this research is available to the public and completely anonymized, there are still many ethical challenges that need to be handled. This includes; Bias, fairness, and responsible use of machine learning [2].

Since survey-based datasets are subject to self-reporting, recall, and non-response biases, this may create problems for individuals who underestimate or overestimate health behaviors (e.g., physical inactivity, alcohol consumption, cigarette smoking), which can have an impact on predicted outcomes from the machine learning model [1]. These issues may disproportionately affect certain populations which may further solidify existing health disparities if not well-monitored or understood.

Class imbalance could lead to bias in the decision-making process of machine learning algorithms due to training on datasets that do not represent all classes of data equally[6]. The model may favor performing higher for the majority population and, therefore, perform poorly for the underrepresented category. Identifying those in high risk is reasonable for prevention; however, the ethical deployment of machine learning requires transparency about the trade-offs involved.

The purpose of the predictive models developed in this project is to inform public health decision-making. They are not intended to be standalone diagnostic tests but should be considered as decision-support systems to augment clinical judgment and policy determinations. Future work will need to include fairness-aware performance evaluation measures and continuous monitoring to ensure equitable and ethical usage of the models.

XIII. Limitations and Future Work

Although an overall strong performance has been presented for this study, there are limitations to consider. First, due to the class imbalance in the data, lower recall scores associated with the low-risk group indicate that high-risk detection is prioritized over low-risk identification, which limits the potential for balanced risk classifications. While this trade-off is considered appropriate for certain screening applications, the model's ability to provide balanced risk classifications is limited.

Second, the BRFSS dataset was self-reported, resulting in noise and uncertainty related to the model's measurements, leading to potential unreliability and reduced generalizability [1]. In addition, due to the potential for data leakage from preprocessing steps, we have taken a conservative approach to feature engineering which may have limited the models' predictive capacities.

Future studies will attempt to overcome these obstacles by utilizing class balancing methods (e.g., SMOTE, cost-sensitive learning), exploring advanced ensemble algorithms (e.g., XGBoost, LightGBM), and implementing explainable AI techniques (e.g., SHAP values) to increase model transparency and interpretability, which are both preconditions for the adoption of these models in healthcare applications.

XIV. Conclusion

The current research presents a comprehensive and secure machine-learning approach for predicting health risk classification through the use of BRFSS survey data. Through extensive data preprocessing, exploratory data analysis, and systematic model assessment, Random Forest was determined to be the optimal model. This model demonstrated high generalization capabilities, as well as excellent sensitivity for identifying individuals at greater risk for poor health.

This work demonstrates the possibility that machine learning may support scalable and evidence-based public health risk-assessment systems. By emphasizing a methodologically rigorous and ethically applicable framework for real-world application, this research sets a strong basis for future efforts in the fields of predictive analytics and population health management.

References

[1] S. Salvi, G. Vu, V. Gurupur, and C. King, "Classifying tooth loss and assessing risk factors in U.S. adults: A machine learning analysis of BRFSS 2022 data," *Electronics*, vol. 14, no. 17, p. 3559, 2025.

[2] Z. Thakkar, Y. Wu, M. Khan, X. Qi, G. A. Hung, N. Kikuta, A. Jamal, M. Srinivasan, R. J. Huang, K. Kim, and G. Kim, "Evaluating the reliability and robustness of racial and ethnic health disparities in cardiometabolic disease in NHANES, NHIS, and BRFSS (2015–2021)," *Journal of the American Heart Association*, vol. 14, no. 5, p. e040029, 2025.

[3] F. Waheed, N. Ehsan, R. Nasir, W. A. Khan, M. F. Khokhar, L. Shahzad, A. Tariq, H. Afzal, and Q. uz Zaman, "Geo-spatial distribution of air pollutants in urban areas and its potential health risk analysis solutions," *Urban Climate*, vol. 61, p. 102380, 2025.

[4] R. Cong, S. Nishimura, A. Ogihara, and Q. Jin, "An exploratory and interpretable approach to estimating latent health risk factors without using domain knowledge," *Big Data Mining and Analytics*, vol. 8, no. 2, pp. 447–457, 2025.

[5] L. Rolle-Lake and E. Robbins, "Behavioral Risk Factor Surveillance System," in *StatPearls [Internet]*, Treasure Island, FL, USA: StatPearls Publishing, 2025.

[6] A. Bouras, "Impact of informal caregiving on health outcomes: A population-based analysis using BRFSS data (2015–2020)," *medRxiv*, preprint, Aug. 2025.

[7] CDC Behavioral Risk Factor Surveillance System (BRFSS) 2015–2022, "Behavioral Risk Factor Surveillance System (BRFSS) data," Kaggle, 2025. [Online]. Available: https://www.kaggle.com/datasets/cdc/behavioral-risk-factor-surveillance-system/data