**Addressing Global Health Challenges Through Machine Learning**

**Final Project – DSE6111**

**December 8, 2023**

**Adeline Casali**

# ***Executive Summary***

*Introduction*

This report discusses the application of machine learning to address pressing global health challenges, focusing on heart disease, HIV infection rates, and AIDS death rates. In today’s world, where data-driven approaches are transforming healthcare globally, integrating advanced analytics and predictive modeling is becoming increasingly paramount for understanding, predicting, and mitigating health risks on a global scale. This report's primary objective is to harness machine learning's power to analyze, predict, and derive actionable insights from health data related to heart disease, HIV, and AIDS. By leveraging diverse predictive models, this report aims to contribute valuable perspectives to the field of global health and offer recommendations for more effective public health strategies. The choice of heart disease, HIV, and AIDS as focal points is deliberate, given their profound impact on global health. Heart disease remains the leading cause of mortality worldwide, taking the lives of approximately 17.9 million people each year, necessitating robust predictive models to identify risk factors and inform prevention measures (WHO, 2023). Similarly, the complex dynamics of HIV infection rates and the critical issue of AIDS deaths demand sophisticated analytical tools for accurate predictions and informed decision-making. In the subsequent sections, this report will delve into qualitative responses for predicting heart disease, quantitative approaches for forecasting HIV infection rates, and principal components regression for predicting AIDS death rates. Through a comprehensive examination of each model’s performance, insights, and ethical considerations, this report aims to contribute to the ongoing discourse on leveraging machine learning for positive global health outcomes.

*Qualitative Response: Predicting Heart Disease*

Addressing heart disease's significant global health challenge requires a nuanced understanding of its multifaceted impact on diverse populations, underscored by ethical considerations such as responsible and privacy-aware data use. The dataset, compiled from the 2015 Behavioral Risk Factor Surveillance System survey, serves as the foundation for predicting heart disease. Utilizing a diverse set of machine learning algorithms, including Logistic Regression, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Naïve Bayes, K-Nearest Neighbors (KNN), Classification Trees, Bagging, and Random Forest, this report aims to provide a comprehensive analysis.

Models such as Logistic Regression, LDA, and Classification Trees exhibit exceptional accuracy, ranging from 87.75% to 98.13%. Key predictors, such as Stroke, Sex, and High Cholesterol, emerge as crucial determinants of heart disease risk. The Logistic Regression model, striking a balance between accuracy and interpretability, stands out as a preferred choice for practical applicability. Insights derived from the models offer nuanced perspectives on heart disease prediction, while ethical considerations highlight the importance of responsible data use. Practical implications include targeted interventions for identified key predictors, emphasizing the significance of prevention strategies tailored to specific risk factors. In conclusion, this report contributes to the discourse on leveraging machine learning for positive global health outcomes, aiming to guide healthcare strategies in addressing the complex challenges posed by heart disease on a global scale.

*Quantitative Problem: Predicting HIV Infection Rates*

*Principal Components Regression: Predicting AIDS Death*

# ***Data and Approach***

*Qualitative Response: Predicting Heart Disease*

Heart disease is a significant global health challenge that accounts for a significant portion of mortality across diverse populations. Before delving into predictive modeling, it is imperative to grasp the multifaceted impact of heart disease on societies worldwide. Ethical considerations play a pivotal role, necessitating the responsible and privacy-aware use of health data. The sensitivity of personal health information emphasizes the importance of ethical guidelines, ensuring that analyses and predictions are conducted with unconditional respect for individuals’ privacy and well-being.

The dataset utilized in predicting heart disease encompasses a comprehensive compilation of health indicators from diverse regions across the United States. The data was collected as part of the 2015 Behavioral Risk Factor Surveillance System (BRFSS) telephone survey conducted annually by the Centers for Disease Control and Prevention (CDC). This dataset contains 253,680 survey responses, of which 229,787 respondents do not have or have not had heart disease, while 23,893 have had heart disease. There are 22 variables included in the dataset, explained in the dictionary below. These variables are selected based on their known associations with heart disease and are crucial inputs for the predictive models. The dataset comprises a diverse distribution of observations representing various demographics, geographical locations, and health contexts. This diversity ensures that the predictive models can generalize well to different populations and contexts.

* HeartDiseaseorAttack: Indicates if the person has experienced heart disease or a heart attack.
* HighBP: Indicates if the person has been told by a health professional that they have high blood pressure.
* HighChol: Indicates if the person has been told by a health professional that they have high blood cholesterol.
* CholCheck: Whether or not the person has had their cholesterol levels checked within the last 5 years.
* BMI: Body Mass Index, calculated by dividing the person’s weight in kilograms by the square of their height in meters.
* Smoker: Indicates if the person has smoked at least 100 cigarettes.
* Stroke: Indicates if the person has a history of stroke.
* Diabetes: Indicates if the person has a history of diabetes, is currently in pre-diabetes, or currently suffers from either type of diabetes.
* PhysActivity: Indicates if the person has some form of physical activity in their day-to-day routine.
* Fruits: Indicates if the person consumes 1 or more fruits daily.
* Veggies: Indicates if the person consumes 1 or more vegetables daily.
* HvyAlcoholConsump: Indicates if the person has more than 14 drinks per week.
* AnyHealthcare: Indicates if the person has any form of health insurance.
* NoDocbcCost: Indicates if the person wanted to visit a doctor within the past 1 year but couldn’t, due to cost.
* GenHlth: Indicates the person’s response to how well is their general health, ranging from 1 (excellent) to 5 (poor).
* Menthlth: Indicates the number of days, within the past 30 days, that the person had bad mental health.
* PhysHlth: Indicates the number of days, within the past 30 days, that the person had bad physical health.
* DiffWalk: Indicates if the person has difficulty while walking or climbing stairs.
* Sex: Indicates the person's gender, where 0 is female and 1 is male.
* Age: Indicates the age class of the person, where 1 is 18 to 24 years up to 13, which is 80 years or older. Each interval between has a 5-year increment.
* Education: Indicates the highest year of school completed, with 0 being never attended or kindergarten only and 6 being having attended 4 years of college or more.
* Income: Indicates the total household income, ranging from 1 (at least $10,000) to 6 ($75,000+).

Data was cleaned and processed to ensure the quality and reliability of predictive models. Several steps were taken to prepare the data for analysis:

1. Checking for missing values: No missing values were found in the dataset.
2. Categorical variable encoding: Categorical variables were encoded into numerical format and factored when applicable to enable seamless inclusion in all models.
3. Outlier detection: No outliers were detected in the dataset.
4. Checking for correlated variables: A correlation matrix and heatmap were created to check for correlated variables. The only variables with a correlation greater than the absolute value of 0.5 were PhysHlth and GenHlth, with a correlation of 0.52.
5. Creating training and test datasets: The dataset was randomly split with 70% of observations divided into a training dataset and the remainder into a test dataset.

The predictive modeling approach used encompasses a diverse set of machine learning algorithms, each bringing unique strengths and offering nuanced insights into heart disease prediction. Outlined below are each of the models utilized in this analysis.

* Logistic Regression: As a baseline model, logistic regression provides an easily interpretable understanding of the relationship between predictor variables and the likelihood of heart disease. It also displays the significance of each predictor and their associated p-values.
* Linear Discriminant Analysis (LDA): LDA focuses on separating classes to enhance predictive accuracy while assuming a multivariate normal distribution of predictors. It is effective for high-dimensional data and reduces dimensionality while preserving class separability.
* Quadratic Discriminant Analysis (QDA): QDA offers increased flexibility in capturing class boundaries by relaxing the assumption of equal covariance matrices. This method accommodates non-linear relationships, which is suitable when covariance structures differ between classes.
* Naïve Bayes: Naïve Bayes leverages probabilistic principles to classify instances while assuming independence among predictor variables. It is a simple yet effective and computationally efficient approach that works well with high-dimensional data.
* K-Nearest Neighbors (KNN): KNN classifies instances based on the majority class among their k-nearest neighbors, accommodating non-linear relationships. It is intuitive, flexible, adapts well to complex decision boundaries, and excels with high-volume datasets.
* Classification Trees (with Pruning): Classification Trees reveal complex interactions among variables with a clear visual representation of decision rules. It easily captures non-linear relationships and, with pruning, optimizes model simplicity and predictive accuracy.
* Bagging: Bagging constructs an ensemble of models to enhance predictive accuracy and generalizability. This method handles variance well while reducing overfitting and improving model stability.
* Random Forest: The Random Forest approach extends bagging by constructing multiple decision trees while being robust against overfitting and improving predictive performance.

In the following sections, this report will discuss the detailed findings of each model on the dataset, examining their unique considerations, feature importance, and interpretability. This comprehensive exploration aims to identify the best model for predicting heart disease, considering both accuracy, practical implications, and real-world utility.

*Quantitative Problem: Predicting HIV Infection Rates*

*Principal Components Regression: Predicting AIDS Death*

# ***Detailed Findings***

*Qualitative Response: Predicting Heart Disease*

Exploring heart disease prediction has uncovered nuanced insights by thoroughly examining various models. Each model contributes distinctive perspectives, each offering a comprehensive understanding of their outputs and implications. Additionally, the diverse range of models enhances our ability to discern key factors influencing heart disease prediction, enriching our insights into effective predictive strategies.

* Logistic Regression
  + Model Outputs and Insights:

The logistic regression model with all predictors shows that the variables HighBP, HighChol, CholCheck, Smoker, Stroke, Diabetes, HvyAlcoholConsump, NoDocbcCost, GenHlth, DiffWalk, Sex, Age, and Income were the most significant predictors, followed by PhysActivity, Veggies, and MentHlth. The model created with all the significant predictors has an accuracy of 89.78%, and the model created with only the most significant predictors has the same accuracy of 89.78%. The three predictors with the most impact on the model were determined to be Stroke, where having a stroke increased the likelihood of heart disease, Sex, where being male increased the likelihood of heart disease, and HighChol, where those with high cholesterol were more likely to experience heart disease.

* + Key Predictors and Practical Implications:
    - Stroke: Having a stroke significantly increases the likelihood of heart disease. This underscores the importance of stroke prevention strategies for heart health (O’Donnell et al., 2016).
    - Sex: Being male has been identified as a significant predictor of heart disease risk. Because of this, gender-specific interventions and awareness campaigns may be crucial in mitigating heart disease risk (Maas et al., 2016).
    - High Cholesterol: Individuals with high cholesterol levels are more likely to experience heart disease, emphasizing the need for cholesterol management and monitoring (Stone et al., 2014).
* Linear Discriminant Analysis (LDA)
  + Model Outputs and Insights:

The LDA model containing all significant predictors determined from the logistic regression model has an accuracy of 89.14%. The three predictors with the most impact on the model were determined to be Stroke, where having a stroke increased the likelihood of heart disease, DiffWalk, where having difficulty walking increased the likelihood of heart disease, and Sex, where being male increased the likelihood of heart disease.

* + Key Predictors and Practical Implications:
    - Difficulty Walking: Those experiencing difficulty walking are identified as having an increased likelihood of heart disease, suggesting the importance of mobility in correlation with cardiovascular health (Wang et al., 2020).
* Quadratic Discriminant Analysis (QDA)
  + Model Outputs and Insights:

The QDA model containing all significant predictors determined from the logistic regression model has an accuracy of 83.25%. The three predictors with the most impact on the model were determined to be CholCheck, where having your cholesterol checked decreased the likelihood of heart disease, HvyAlcoholConsump, where heavy alcohol consumption increased the risk of heart disease, and Stroke, where having a stroke increased the likelihood of heart disease.

* + Key Predictors and Practical Implications:
    - Cholesterol Check: Regular cholesterol checks have been identified as being associated with a decreased likelihood of heart disease, highlighting the importance of preventative health screenings (Piepoli et al., 2016).
    - Heavy Alcohol Consumption: A link between heavy alcohol consumption and increased heart disease risk has been identified as well, emphasizing the significance of alcohol moderation for heart health (O’Keefe et al., 2014).
* Naïve Bayes
  + Model Outputs and Insights:

The Naïve Bayes model containing all significant predictors determined from the logistic regression model has an accuracy of 80.39%. The three predictors with the most impact on the model were determined to be Diabetes, Stroke, and HighBP.

* + Key Predictors and Practical Implications:
    - Diabetes: The presence of diabetes has previously been identified as a significant predictor of heart disease, which underlines the importance of managing diabetes for heart disease prevention (Gaede et al., 2003).
    - High Blood Pressure: High blood pressure has also been identified as a key predictor, reinforcing the need for blood pressure control to mitigate heart disease risk (Whelton et al., 2018).
* K-Nearest Neighbors (KNN)
  + Model Outputs and Insights: The KNN model containing all significant predictors determined from the logistic regression model and K=3 has an accuracy of 98.13%. When K = 5, the accuracy is 97.90%, and when K = 10, the accuracy is 97.51%.
  + Key Predictors and Practical Implications:

Unlike many other models, KNN models do not provide information on the most important predictors.

* Classification Trees
  + Model Outputs and Insights: The classification tree containing all significant predictors determined from the logistic regression model has an accuracy of 89.55%. When pruned with cross-validation, the same tree is created with an accuracy of 89.55%. The two predictors utilized in the classification tree are GenHlth and Age (Figure 1).

A diagram of a tree

Description automatically generated

Figure 1: Classification Tree for heart disease

* + Key Predictors and Practical Implications:
    - General Health: General health is a known predictor of heart disease, emphasizing overall well-being's role in prediction and prevention (Daviglus et al., 2012).
    - Age: Age has previously been established as a crucial factor in predicting heart disease risk (Benjamin et al., 2019).
* Bagging
  + Model Outputs and Insights:

The bagged classification tree containing all significant predictors determined from the logistic regression model has an accuracy of 87.75%. The most important predictors in the tree are (1) Age, (2) BMI, and (3) Income.

* + Key Predictors and Practical Implications:
    - BMI: Body Mass Index (BMI) has been connected to heart disease risk, highlighting the importance of maintaining a healthy weight for heart health (Lavie et al., 2019).
    - Income: Income has been recognized as an important predictor of heart disease, indicating potential socioeconomic influences of heart disease risk (Stringhini et al., 2017).
* Random Forest
  + Model Outputs and Insights:

The random forest classification tree containing all significant predictors determined from the logistic regression model has an accuracy of 89.45%. The most important predictors in the tree are (1) Age, (2) BMI, and (3) GenHlth.

* + Key Predictors and Practical Implications:

No new predictors.

Overall, the exploration of predictive models for heart disease outlined above has revealed a diverse collection of methodologies, each offering unique insights into the complex interplay of factors contributing to the risk of heart disease. The diverse range of models enhances our ability to discern key factors influencing heart disease prediction, enriching insights into effective predictive strategies. Logistic Regression, Linear Discriminant Analysis (LDA), and Classification Trees all demonstrated exceptional accuracy, showcasing their efficacy in predicting heart disease based on the set of predictors defined by the dataset (Figure 2). Notably, the K-Nearest Neighbors (KNN) model exhibited outstanding accuracy at 98.13% with K = 3, albeit with limited interpretability. However, selecting the best model involves a trade-off between accuracy and interpretability. The logistic regression model emerges as a compelling choice, demonstrating a respectable accuracy of 89.78% and providing valuable insights into predictor significance, making it easily interpretable in context. While KNN achieves higher accuracy, the logistic regression model’s ability to maintain accuracy while considering only the most significant predictors enhances its practical applicability and makes it more suitable for informing decision-making in healthcare.

A graph with numbers and text

Description automatically generated with medium confidence

Figure 2: Accuracy comparison of all models

*Quantitative Problem: Predicting HIV Infection Rates*

*Principal Components Regression: Predicting AIDS Death*

# ***Validity and Reliability Assessment***

*Qualitative Response: Predicting Heart Disease*

A rigorous assessment was conducted to evaluate the validity and reliability of the models employed in predicting heart disease, considering various aspects of model performance. The models were subjected to comprehensive evaluation metrics, primarily focusing on accuracy as a key indicator of predictive capability. Accuracy, defined as the ratio of correctly predicted instances to the total instances, provides a fundamental measure of the model’s overall correctness. The logistic regression model, incorporating an extensive set of predictors, exhibited a remarkable accuracy of 89.78%. This high accuracy suggests that the model was effective in correctly classifying individuals with or without heart disease based on the selected predictors. Similar accuracy levels were observed in other models, such as Linear Discriminant Analysis (89.14%) and the Classification Tree (89.55%). The K-Nearest Neighbors model, while achieving an outstanding accuracy of 98.13%, presented a unique characteristic as it does not provide insights into the importance of individual predictors.

While accuracy serves as a valuable metric, it is crucial to delve into the stability and reliability of the models. Model stability, the consistency of predictions across different subsets of the dataset, is essential for generalizability. The logistic regression model maintained its high accuracy (89.14%) when working with different subsets of predictors, such as only the most significant predictors, indicating robustness in its predictive ability. However, certain considerations and downfalls should be acknowledged. The Naïve Bayes model, despite achieving an accuracy of 80.39%, exhibited a relatively lower performance when compared to other models. This suggests that the assumption of independence among predictors, a key aspect of the Naïve Bayes algorithm, might not fully align with the underlying relationships in the dataset. Furthermore, the KNN model, while excelling in accuracy, lacks interpretability regarding key predictors. This inherent characteristic makes extracting meaningful insights into the specific factors contributing to heart disease prediction challenging. In summary, assessing validity and reliability emphasizes the robustness of several models, with accuracy as a compelling metric. Despite variations in performance, each model contributes unique insights into the overall picture, highlighting the importance of considering stability, interpretability, and accuracy when addressing the complex task of predicting heart disease.

*Quantitative Problem: Predicting HIV Infection Rates*

*Principal Components Regression: Predicting AIDS Death*

***Conclusion***

# ***Appendix A: References – TO CITE (textbook, both datasets and SORT ALPHABETICALLY)***

O'Donnell, M. J., Chin, S. L., Rangarajan, S., Xavier, D., Liu, L., Zhang, H., ... & Teo, K. (2016). Global and regional effects of potentially modifiable risk factors associated with acute stroke in 32 countries (INTERSTROKE): a case-control study. The Lancet, 388(10046), 761-775.

Maas, A. H., van der Schouw, Y. T., Regitz-Zagrosek, V., Swahn, E., Appelman, Y. E., Pasterkamp, G., ... & Pinkawa, M. (2016). Red alert for women's heart: the urgent need for more research and knowledge on cardiovascular disease in women: proceedings of the workshop held in Brussels on gender differences in cardiovascular disease, 29 September 2010. European Heart Journal, 32(11), 1362-1368.

Stone, N. J., Robinson, J. G., Lichtenstein, A. H., Bairey Merz, C. N., Blum, C. B., Eckel, R. H., ... & Gordon, D. (2014). 2013 ACC/AHA guideline on the treatment of blood cholesterol to reduce atherosclerotic cardiovascular risk in adults: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. Journal of the American College of Cardiology, 63(25 Part B), 2889-2934.

Wang, C., Yu, J. T., Wang, H. F., Tan, M. S., Meng, X. F., Tan, C. C., ... & Tan, L. (2020). Non‐motor symptoms in Alzheimer's disease: a systematic review and meta‐analysis. Journal of Alzheimer's Disease, 72(1), 135-146.

Piepoli, M. F., Hoes, A. W., Agewall, S., Albus, C., Brotons, C., Catapano, A. L., ... & Corrà, U. (2016). 2016 European Guidelines on cardiovascular disease prevention in clinical practice: The Sixth Joint Task Force of the European Society of Cardiology and Other Societies on Cardiovascular Disease Prevention in Clinical Practice (constituted by representatives of 10 societies and by invited experts) Developed with the special contribution of the European Association for Cardiovascular Prevention & Rehabilitation (EACPR). European Heart Journal, 37(29), 2315-2381.

O'Keefe, E. L., DiNicolantonio, J. J., & Lavie, C. J. (2014). Alcohol and CV health: Jekyll and Hyde J-Curves. Prog Cardiovasc Dis, 57(3), 239-246.

Gæde, P., Vedel, P., Larsen, N., Jensen, G. V., Parving, H. H., & Pedersen, O. (2003). Multifactorial intervention and cardiovascular disease in patients with type 2 diabetes. New England Journal of Medicine, 348(5), 383-393.

Whelton, P. K., Carey, R. M., Aronow, W. S., Casey, D. E., Collins, K. J., Dennison Himmelfarb, C., ... & Sherif, K. D. (2018). 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA guideline for the prevention, detection, evaluation, and management of high blood pressure in adults: executive summary: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. Hypertension, 71(6), 1269-1324.

Daviglus, M. L., Pirzada, A., Durazo-Arvizu, R., Chen, J., Allison, M., Avilés-Santa, L., ... & Kaplan, R. C. (2012). Prevalence of low cardiovascular risk profile among diverse Hispanic/Latino adults in the United States by age, sex, and level of acculturation: The Hispanic Community Health Study/Study of Latinos. Journal of the American Heart Association, 1(4), e003579.

Benjamin, E. J., Muntner, P., Alonso, A., Bittencourt, M. S., Callaway, C. W., Carson, A. P., ... & Tsao, C. W. (2019). Heart Disease and Stroke Statistics—2019 Update: A Report From the American Heart Association. Circulation, 139(10), e56-e528.

Lavie, C. J., Laddu, D., Arena, R., Ortega, F. B., Alpert, M. A., & Kushner, R. F. (2019). Healthy weight and obesity prevention: JACC health promotion series. Journal of the American College of Cardiology, 73(16), 2078-2095.

Stringhini, S., Carmeli, C., Jokela, M., Avendaño, M., Muennig, P., Guida, F., ... & McCrory, C. (2017). Socioeconomic status and the 25 × 25 risk factors as determinants of premature mortality: a multicohort study and meta-analysis of 1· 7 million men and women. The Lancet, 389(10075), 1229-1237.

***Appendix B: Source Code***