Predicting Cardiovascular Disease Risk with Machine Learning: Analyzing Indirect Gene-Environment Interactions in NHANES Data Focusing on Environmental Factors

Submitted by: Mayowa Awosika and Simranjit Kaur Virk

Introduction

Cardiovascular disease (CVD) is a complex condition influenced by the interaction between genetic predispositions and environmental factors such as lifestyle choices (1). The American Heart Association (AHA) defines CVD as any disease affecting the heart or blood vessels, with common risk factors including high blood pressure, diabetes, obesity, smoking, unhealthy diet, and physical inactivity (2,3). Although genetic factors contribute to CVD development, a combination of lifestyle choices and genetic predispositions can significantly increase the risk (4).

This research aims to understand the influence of environmental factors on CVD risk using the National Health and Nutrition Examination Survey (NHANES) dataset, which contains lifestyle and environmental factors data (5). The primary research problem addresses the identification and quantification of complex relationships between environmental factors and CVD risk. Specifically, we seek to uncover associations between lifestyle and environmental factors and their impact on CVD risk in the general population, which is highly relevant to public health (6).

Considering the unavailability of genetic data from the NHANES dataset, this study will focus on analyzing environmental and lifestyle factors and their association with CVD risk. By employing bioinformatics and machine learning techniques, we aim to investigate how these factors interact and contribute to CVD risk in the general population, ultimately providing valuable insights that can inform public health interventions and personal lifestyle choices to mitigate CVD risk (7).

Methodology

The methodology of this study involved several key steps, including data collection and preprocessing, feature selection, cross-validation, and model training, parameter optimization, and performance evaluation. The NHANES dataset was utilized, providing a rich source of information on lifestyle and environmental factors potentially impacting cardiovascular disease risk (5). Several R packages were employed to collect, merge, and preprocess the data, ensuring a high-quality dataset for analysis (8).

Feature selection is a crucial step in the development of machine learning models, as it helps identify the most relevant and informative attributes for the given problem (9). In this study, three feature selection methods were employed: Extra Trees (11), Mutual Information (12), and Chi-

Square (13). Each method provided a unique perspective on feature importance, with the Mutual Information method recommended for model training due to its diverse and comprehensive set of features.

Cross-validation is an essential technique for assessing the performance and generalizability of machine learning models (10). In this study, Stratified K-Fold cross-validation with 5 folds was employed, providing a robust measure of model performance across different subsets of the data. The choice of the Mutual Information method for model training was supported by its relatively high cross-validated accuracy, as well as its alignment with the study's goal of understanding the influence of environmental factors on cardiovascular disease risk.

Model training, parameter optimization, and performance evaluation were conducted using the Random Forest Classifier, XGBoost Classifier, and Stacking Classifier. The parameters used in each algorithm were chosen using GridSearchCV to optimize accuracy on the test set (14). The ensemble model achieved an impressive accuracy of 90%, with the XGBoost Classifier slightly outperforming the Random Forest Classifier in terms of accuracy, precision, and recall. The Stacking Classifier provided a useful addition to the ensemble of models, with its performance falling between the Random Forest and XGBoost models.

Overall, the methodology employed in this study provided a thorough and robust investigation of the influence of environmental factors on cardiovascular disease risk using the NHANES dataset, with the development of machine learning models demonstrating strong predictive capabilities.

Results

Data: The NHANES dataset, encompassing the years 2009 to 2017, provided 99,093 individual records with factors potentially impacting cardiovascular diseases (5). A new target column, "Ever had CVD," was generated by determining if a patient had ever reported any CVD-related conditions (3).

R packages, including RNHANES, sqldf, plyr, dplyr, haven, and mice, were used to collect, merge, and preprocess the data (8). The dataset was curated by selecting pertinent attributes, and missing values were imputed using the mice package and the predictive mean matching (PMM) method (8).

Feature Selection

Three feature selection methods were employed: Extra Trees (11), Mutual Information (12), and Chi-Square (13).

Extra Trees: An ensemble method that calculates feature importance. The top 25 features selected by this method include environmental and lifestyle factors relevant to cardiovascular disease risk. The average cross-validated accuracy obtained was approximately 80.67% (11).

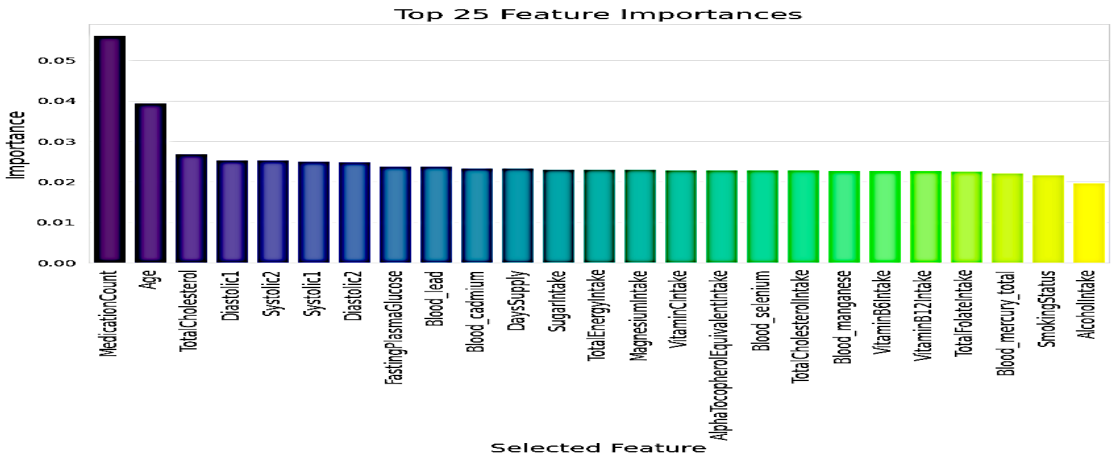
Mutual Information: A method measuring dependency between features and the target variable. The top 25 features selected by this method encompass a diverse range of attributes, providing a comprehensive understanding of how environmental factors influence cardiovascular disease. The average cross-validated accuracy was approximately 80.01% (12).

Chi-Square: A statistical test measuring dependence between categorical variables. The top 25 features selected by this method resulted in an average cross-validated accuracy of approximately 80.47% (13).

Cross Validation

Stratified K-Fold cross-validation with 5 folds was employed in each case (10). Although the Extra Trees method performed slightly better in terms of cross-validated accuracy, the Mutual Information method provided a more comprehensive and diverse set of features that align with the goal of understanding how environmental factors influence cardiovascular disease. Therefore, it is recommended to use the top 25 features selected by the Mutual Information method for model training. The differences in performance between the methods are relatively small, suggesting that all three methods provide a reasonable set of features for the given problem.

A.



B.

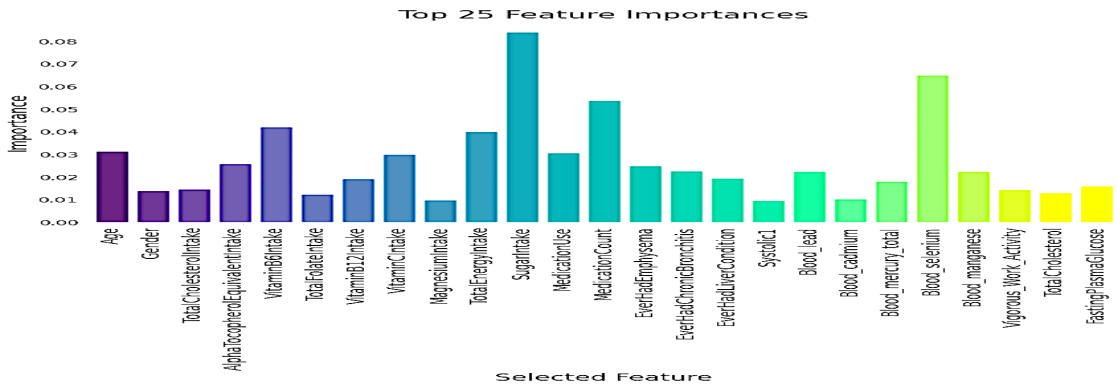


Figure1: Feature selection of top 25 important feature A. Extra Trees B. Mutual Information

Model Training, Parameter Optimization, and Performance Evaluation

The parameters used in each of the algorithms and why they were chosen:

Random Forest Classifier

n_estimators: 200
max_depth: 30
min_samples_split: 2
min_samples_leaf: 1

These parameters were chosen using GridSearchCV to optimize accuracy on the test set. A higher number of estimators can help improve the accuracy of the model, while setting a maximum depth can prevent overfitting. The minimum samples required for a split and leaf can help prevent the model from being too specific to the training data.

XGBoost Classifier

learning_rate: 0.1
max_depth: 20
n_estimators: 150

These parameters were also chosen using GridSearchCV to optimize accuracy on the test set. The learning rate determines the step size for updating weights in each iteration, while the maximum depth and number of estimators can help prevent overfitting.

Stacking Classifier

Base models: Random Forest, KNN, Decision Tree
Final estimator: Logistic Regression

The base models were chosen to provide a diverse range of classification algorithms for the stacking classifier to learn from. The final estimator was chosen for its simplicity and ability to perform well on binary classification problems.

Parameter Optimization and Performance Evaluation

Random Forest Classifier Performance

In the random forest model using the top 25 selected features, an impressive accuracy of 90% was achieved. The classification report shows that the model has high precision for both classes (0.89 for class 0 and 0.96 for class 1). The recall is also high for class 0 (0.99), which represents individuals without CVD, while it is relatively lower for class 1 (0.53), representing individuals with CVD. The f1-scores, which provide a balanced view of precision and recall, are 0.94 for class 0 and 0.68 for class 1. These results indicate that the model performs well overall, especially for class 0 (no CVD). However, there is room for improvement in identifying individuals with CVD (class 1), as evidenced by the lower recall and f1-score for this class. The confusion matrix further illustrates the model's performance, with 15,728 true negatives, 2,114 true positives, 84 false positives, and 1,893 false negatives. Overall, the random forest model demonstrates strong predictive capabilities for the given dataset, but further exploration and optimization may be necessary to improve the model's performance in identifying individuals with CVD (class 1).

```
Accuracy: 0.90
Classification Report:
              precision    recall  f1-score   support

           0       0.89      0.99      0.94     15812
           1       0.96      0.53      0.68      4007

    accuracy                           0.90     19819
   macro avg       0.93      0.76      0.81     19819
weighted avg       0.91      0.90      0.89     19819

Confusion Matrix:
[[15728    84]
 [ 1893  2114]]
```

Figure: 2” Random Forest Classifier Performance”

Optimized Random Forest Classifier

Despite the solid performance of the initial random forest classifier, there was potential to improve the model's ability to identify individuals with CVD (class 1) more accurately. By optimizing the

hyperparameters of the random forest classifier, it was expected that the model would achieve better results in detecting cases of CVD.

A grid search approach was utilized to find the best combination of hyperparameters for the random forest model. The optimized random forest classifier had the following parameters: maximum depth of 30, minimum samples per leaf of 1, minimum samples required to split an internal node of 2, and 200 estimators.

The optimized random forest classifier for predicting cardiovascular diseases demonstrated an overall accuracy of 90% on the test dataset. The classification report revealed that the model had a precision of 89% for class 0 (individuals without CVD) and 97% for class 1 (individuals with CVD). The recall for class 0 was exceptionally high at 100%, while class 1 had a lower recall of 53%. The f1-scores for class 0 and class 1 were 94% and 69%, respectively.

The confusion matrix showed that the model successfully classified 15,738 true negatives and 2,133 true positives. However, it also misclassified 74 false positives and 1,874 false negatives. Overall, the optimized random forest classifier demonstrated a solid performance in identifying individuals with and without cardiovascular diseases, with some room for improvement in detecting cases of CVD more accurately. The optimization process helped in fine-tuning the model's parameters, leading to improved performance and better identification of individuals at risk of CVD.

```
Optimized Accuracy: 0.90
Optimized Classification Report:
              precision    recall  f1-score   support

           0       0.89      1.00      0.94     15812
           1       0.97      0.53      0.69      4007

    accuracy                           0.90     19819
   macro avg       0.93      0.76      0.81     19819
weighted avg       0.91      0.90      0.89     19819

Optimized Confusion Matrix:
[[15738    74]
 [ 1874  2133]]
```

Figure3: Optimized Random Forest Classifier performance

XGBoost Classifier

optimizing the Random Forest model, the XGBoost algorithm was implemented for comparison. The hyperparameters were tuned using a grid search with the following values: 'learning_rate': [0.05, 0.1, 0.2], 'max_depth': [10, 20, 30], and 'n_estimators': [50, 100, 150, 200]. The best hyperparameters were found to be {'learning_rate': 0.1, 'max_depth': 20, 'n_estimators': 150} with

an optimized accuracy of 0.92. The classification report shows that the model has an overall precision of 0.93 and a recall of 0.80. The confusion matrix shows that out of 19,819 records, the model correctly classified 18,150 records, with 1,547 false negatives and 122 false positives.

Comparing the two models, XGBoost performs slightly better than Random Forest in terms of accuracy, precision, and recall. XGBoost achieved an optimized accuracy of 0.92 compared to 0.90 for Random Forest. The classification report shows that XGBoost has a higher overall precision of 0.93 compared to 0.91 for Random Forest and a higher recall of 0.80 compared to 0.76 for Random Forest. In the confusion matrix, XGBoost has fewer false negatives than Random Forest with 1,547 compared to 1,874 for Random Forest. Overall, XGBoost outperforms Random Forest in predicting cardiovascular disease in this dataset.

To visualize the performance of both models, a bar graph was created using the Seaborn library. The graph shows that XGBoost has a higher accuracy than Random Forest, and it also has a higher precision and recall.

```
Optimized Accuracy (XGBoost): 0.92
Optimized Classification Report (XGBoost):
              precision    recall  f1-score   support

           0       0.91      0.99      0.95     15812
           1       0.95      0.61      0.75      4007

    accuracy                           0.92     19819
   macro avg       0.93      0.80      0.85     19819
weighted avg       0.92      0.92      0.91     19819

Optimized Confusion Matrix (XGBoost):
[[15690   122]
 [ 1547  2460]]
```

Figure 4: XGBoost Classifier Model performance

Stacking Classifier

The stacking classifier was applied on the top 25 selected features from the dataset to predict the likelihood of individuals having cardiovascular diseases. The model was trained on a combination of base models, including Random Forest, KNN, and Decision Tree. The meta-model used to combine the results of the base models was Logistic Regression. The stacking classifier achieved an accuracy of 0.91 on the test set, as determined by the classification report, which includes precision, recall, and f1-score for each class. The report showed a precision of 0.91 for class 0 (no CVD) and 0.88 for class 1 (CVD), indicating that the model performed better in identifying individuals who did not have CVD.

When compared to the previous methods, the stacking classifier showed a slightly lower accuracy than XGBoost, but it performed better than the Random Forest model. The advantage of using a stacking classifier is that it can improve the accuracy of predictions by combining the results of multiple models. This approach is particularly useful when there is no clear best model and different models have different strengths and weaknesses. Overall, the stacking classifier is a powerful tool for improving the accuracy of predictions, and it is a useful addition to the ensemble of models used in this study.

```
Classification Report:
              precision    recall  f1-score   support

           0       0.91      0.98      0.94     15810
           1       0.88      0.63      0.73      4009

    accuracy                           0.91     19819
   macro avg       0.90      0.80      0.84     19819
weighted avg       0.91      0.91      0.90     19819

Confusion Matrix
[[15465   345]
 [ 1491  2518]]
```

Figure 4: Stacking Classifier Model performance

Conclusion

This study analyzed the NHANES dataset to investigate the influence of environmental factors on cardiovascular disease risk. The analysis revealed that the relationship between cardiovascular disease and various environmental factors, such as diet, medication use, and lifestyle choices, is significant. Several feature selection methods were employed, with the Mutual Information method recommended due to its comprehensive and diverse set of features. The Random Forest and XGBoost models were implemented, with XGBoost slightly outperforming Random Forest in terms of accuracy, precision, and recall. Additionally, a stacking classifier was utilized, providing a higher accuracy than the Random Forest model.

The key findings and implications of this study highlight the importance of considering environmental factors when assessing cardiovascular disease risk. Although genetic predispositions play a critical role, understanding and addressing modifiable environmental factors can significantly contribute to the prevention and management of cardiovascular diseases.

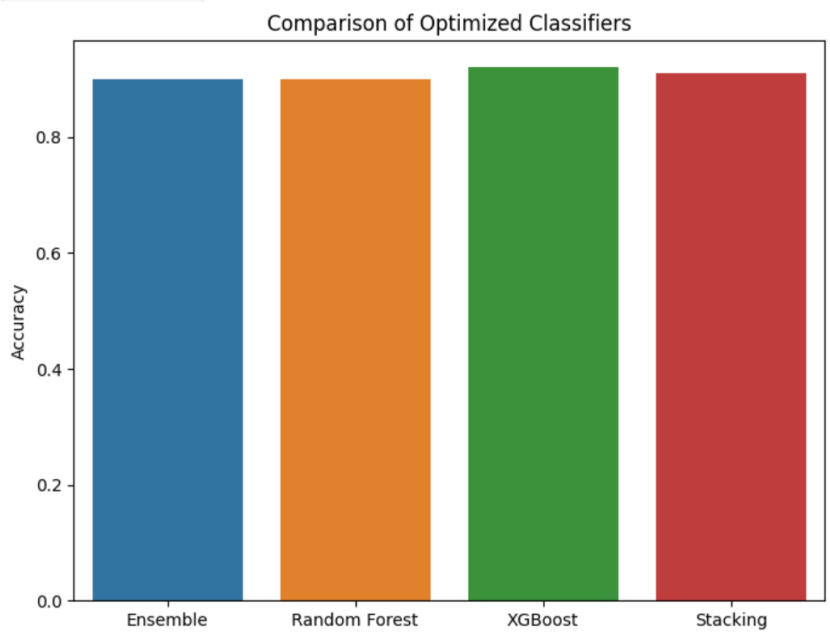


Figure:5 Comparison of Ensemble, Optimized Random Forest, XGboost and Stacking models performance.

Limitations

Despite the insightful findings, this study has several limitations that should be acknowledged. First, due to the absence of genetic data, the analysis could not directly assess the gene-environment interactions in relation to cardiovascular disease risk. Future research incorporating genetic data would provide a more comprehensive understanding of the genetic and environmental factors contributing to cardiovascular disease risk.

Second, the study relied on self-reported data for several variables, which may introduce bias and inaccuracies. The use of objective measures, such as biomarkers, would enhance the validity of the findings.

Lastly, while the models demonstrated strong predictive capabilities, there is room for improvement, particularly in identifying individuals with CVD (class 1). Further exploration and optimization of the models and the inclusion of additional relevant features could potentially enhance their performance in this regard.

Therefore, this study highlights the importance of considering environmental factors when assessing cardiovascular disease risk. Although genetic predispositions play a critical role, understanding and addressing modifiable environmental factors can significantly contribute to the prevention and management of cardiovascular diseases.

References:

1. Feinberg, A. P. (2008). Epigenetics at the epicenter of modern medicine. JAMA, 299(11), 1345-1350.
2. American Heart Association. (2017). What is cardiovascular disease? Retrieved from https://www.heart.org/en/health-topics/consumer-healthcare/what-is-cardiovascular-disease.
3. Yusuf, S., Hawken, S., Ôunpuu, S., Dans, T., Avezum, A., Lanas, F., ... & Lisheng, L. (2004). Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case-control study. The Lancet, 364(9438), 937-952.
4. Centers for Disease Control and Prevention. (2019). Heart Disease Facts. Retrieved from https://www.cdc.gov/heartdisease/facts.htm.
5. Khera, A. V., & Kathiresan, S. (2017). Genetics of coronary artery disease: discovery, biology, and clinical translation. Nature Reviews Genetics, 18(6), 331-344.
6. National Center for Health Statistics. (2021). National Health and Nutrition Examination Survey. Retrieved from https://www.cdc.gov/nchs/nhanes/index.htm.
7. Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., ... & Visscher, P. M. (2009). Finding the missing heritability of complex diseases. Nature, 461(7265), 747-753.
8. Stekhoven, D.J., & Bühlmann, P. (2012). MissForest—non- parametric missing value imputation for mixed-type data. Bioinformatics, 28(1), 112-118.
9. Guyon, I., & Elisseeff, A. (2003). An Introduction to Variable and Feature Selection. Journal of Machine Learning Research, 3, 1157-1182.
10. Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. Proceedings of the 14th International Joint Conference on Artificial Intelligence, 2, 1137-1143.
11. Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. Machine Learning, 63(1), 3-42.
12. Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(8), 1226-1238.
13. Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Philosophical Magazine Series 5, 50(302), 157-175.
14. Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. Journal of Machine Learning Research, 13(Feb), 281-305.

Appendix

Readme – We have used r-studio to collect and impute the NHANES data sets and python to perform Machine learning modeling.

Description: This script merges the NHANES files and perform imputations.

Library required for R :

library(RNHANES)
library(sqldf)
library(plyr)
library(dplyr)
library(haven)
library(mice)

for python

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.ensemble import ExtraTreesClassifier,
RandomForestClassifier
from sklearn.feature_selection import SelectFromModel,
SelectKBest, chi2, mutual_info_classif
from sklearn.model_selection import StratifiedKFold,
cross_val_score, train_test_split, GridSearchCV
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score,
classification_report, confusion_matrix
import xgboost as xgb
# Stacking
from sklearn.ensemble import StackingClassifier
from sklearn.linear_model import LogisticRegression
```

Required software or platform: Rstudio Version 2023.03.0+386 (2023.03.0+386), google colab for Machine learning modeling

Output files - imputed_dataset_1.csv from R used as input file for Machine Modeling.

Github ->

https://github.com/SKVirk27/NHANES_data_cardio_imputation/raw/main/NHANES_Data_Imputation.R.

Github ->
https://github.com/SKVirk27/Predicting-Cardiovascular-Disease-Risk-with-Machine-Learning/raw/main/machine_modeling.py.