Chronic Disease Risk Prediction

Harikrishna Yarraguntla [1] Divya Sai Sundru [2] Navaneeth Gopireddy [3]

University of Michigan Dearborn [1]

IMSE 586

Introduction

Chronic diseases are long-lasting medical conditions that can significantly impact an individual's quality of life and contribute to healthcare costs. These diseases, such as heart disease, diabetes, cancer, and respiratory diseases, can be controlled but not always cured. Preventive measures and early interventions are crucial to manage these diseases effectively. [1] Our project aims to predict chronic disease risk based on age, physical activity, diet, habits, and medical history.

Dataset Description

- We have used the National Health and Nutrition Examination Survey (NHANES) data between Jan 2017 and March 2020.
- NHANES, supervised by National Center for Health Statistics assesses health and nutrition in US adults and children
- The data which is available on CDC portal covers diverse factors contributing to health-related issues among participants
- The dataset comprises records of 15,560 participants over five categories:

Category	Description
Demographic	Includes details such as gender, age, race, country of birth etc.
Dietary	Encompasses data on food consumption, nutrient intake, and vitamin levels
Examination	Consists of information on health metrics like blood pressure and cholesterol
Laboratory	Information on chemical components in the body from medical examinations
Questionnaire	Responses of interview questions to the participants and it serves as a critical component

Table 1. Description of NHANES Data Categories

Research Questions

- What are the optimal methods for effectively cleaning and preparing extensive data for modeling?
- Can lifestyle features of a person can be used to predict the risk of chronic disease?
- Which factors in our daily lives significantly contribute to the development of chronic diseases?
- How can we improve feature selection to capture crucial details for predicting the risk of chronic diseases, including lifestyle features, medical history, and social factors?
- Which machine learning models, such as Logistic Regression, Decision Tree, Support Vector Machine, or K-Neighbors Classifier, are most effective in classifying the risk of specific chronic diseases with maximum accuracy and minimal overfitting?
- Is it feasible to develop a user-friendly dashboard for convenient accessibility to this information?

Methodology

1. Data Processing and Cleaning: Extensive data underwent thorough cleaning and filtering, focusing on relevant variables for the prediction. This involved renaming, recoding, and integrating datasets, resulting in a reduction from 15,560 to 5,314 records and 44 features to ensure integrity by removing missing values and redundant information.
2. Data Visualization: Thorough exploration using various visualizations highlighted chronic disease prevalence patterns and the impact of lifestyle on health parameters. In the below figure, you can clearly observe people who exercise comparably have less cholestrol.

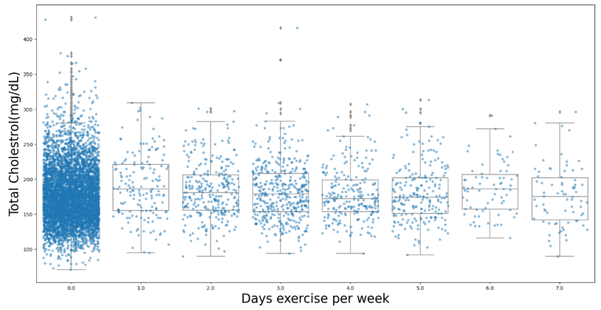

Figure 1. Cholestrol Vs Average Exercise days per week

Correlation analysis aided in refining the feature set by removing 10 strongly inter-correlated variables, reducing the feature set to 32 variables.

3. Model Pre-processing: This step involves splitting data into training and testing sets (80-20 ratio), encoding categorical variables to binary equivalent(0/1) for model interpretation, and standard scaling numerical features. These processes ensure model readiness, prevent bias, and enhance efficiency in training predictive models.
4. Initial Model Building: The training dataset is used for different models namely: Logistic Regression, Decision Tree Classifier, Random Forest Classifier, K-Neighbors Classifier, Support Vector Machine and Neural Networks.
5. Forward Selection: This method sequentially adds variables to build predictive models based on their significance, aiming to reach a streamlined feature set with reduced complexity and better accuracy.
6. Hyper Parameter Tuning: Grid Search optimizes model performance by testing multiple hyperparameter combinations, determining the final parameters with higher accuracy.

Model Evaluation

Model	Accuracy	Recall	F1 Score
Logistic Regression	0.66	0.66	0.56
Decision Tree Classifier	0.49	0.49	0.51
Random Forest Classifier	0.67	0.67	0.54
K-Neighbors Classifier	0.60	0.60	0.54
Support Vector Machine	0.66	0.66	0.53
Neural Networks	0.66	-	-

Table 2. Initial Model Performance metrics

Model	Features Eliminated	Accuracy
Logistic Regression	18	0.66
Decision Tree Classifier	18	0.51
Random Forest Classifier	18	0.66
K-Neighbors Classifier	18	0.60
Support Vector Machine	18	0.66

Table 3. Forward Selection-Model Performance metrics

Model Evaluation

Model	Parameters Tuned	Accuracy
Logistic Regression	C: 0.1, penalty: l2, solver: liblinear	0.67
Decision Tree Classifier	criterion: gini, max_depth: 5, max_features: log2, splitter: random	0.66
Random Forest Classifier	bootstrap: True, criterion: gini, max_features: sqrt, n_estimators: 300	0.66
K-Neighbors Classifier	algorithm: auto, leaf_size: 20, n_neighbors: 9, p: 2, weights: distance	0.64
Support Vector Machine	C: 1, decision_function_shape: ovo, degree: 2, gamma: scale, kernel: rbf	0.66

Table 4. Hyperparameter Tuning-Model Performance metrics

The Logistic Regression model with C: 0.1, penalty: l2, solver: 'liblinear' was chosen as the final model due to its superior accuracy of 67% compared to other models, offering both simplicity and effectiveness.

Dashboard Development

A dashboard is being developed using Dash, utilizing the finalized model. This dashboard takes features ingested in the model as inputs, employs the model to predict the disease outcome, and displays the resulting output.

Figure 2. Snip of the Dashboard being published on localhost

Conclusion

Despite significant challenges posed by reduced dataset size and imbalanced class distribution, the project successfully developed predictive models for chronic disease risks based on lifestyle factors. The efforts in data processing, visualization and model optimization led to the selection of the Logistic Regression model as the final model, offering 67% accuracy. A User-friendly dashboard is also created to access this model. However, the limitations of reduced dataset size and imbalanced class distribution notably impacted model accuracies, falling below 70% across all models tried. These constraints highlight the real-world complexities and emphasize the need for larger, balanced datasets to get better model accuracy in predicting chronic disease risks based on lifestyle and medical history.

[2] [3]

References

[1] Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, 19(6):1236–1246, 2018.

[2] Wullianallur Raghupathi and Viju Raghupathi. Big data analytics in healthcare: promise and potential. *Health information science and systems*, 2:1–10, 2014.

[3] Jingmei Yang, Xinglong Ju, Feng Liu, Onur Asan, Timothy S Church, and Jeff O Smith. Prediction for the risk of multiple chronic conditions among working population in the united states with machine learning models. *IEEE Open Journal of Engineering in Medicine and Biology*, 2:291–298, 2021.