Predictive Modeling for Heart Attack Risk

Mohammad.H Nikkhah,
Amir Hossein Yousefvand,
Ali Akhavan Hosseini

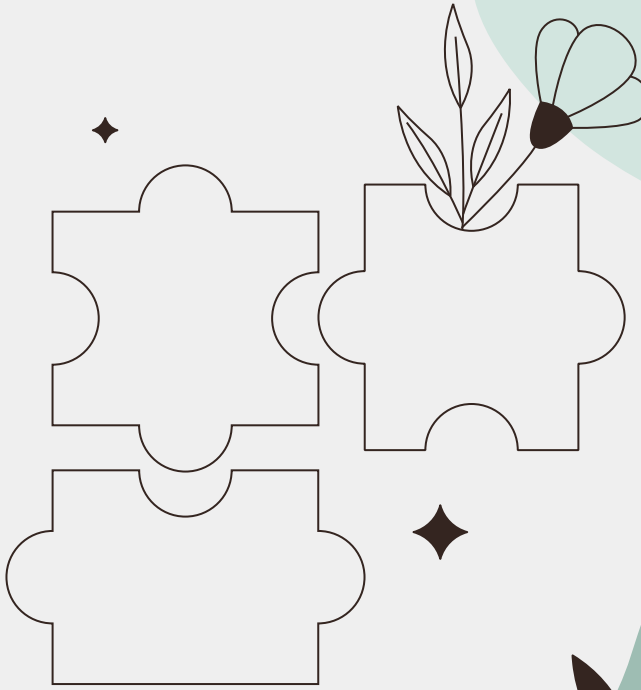


Table of contents

01
Introduction

Cardiovascular diseases (CVDs)

- Leading cause of death globally
- Primary symptoms :
  - Heart attack
  - stroke

19,700,000

Estimated number of lives taken by CVDs each year

Heart Attack Prediction Dataset

Source

www.kaggle.com

Structure

25 features :
13 categorical,
12 ratio

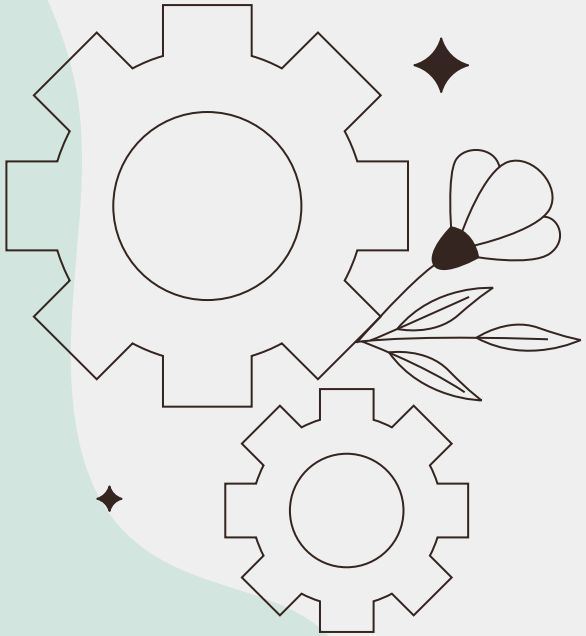
Target

Heart Attack Risk
(yes/ no)

Heart Attack Prediction Dataset

- **Demographic info** : Age, Sex, BMI
- **Habits** : Smoking, Physical activity per day, etc.
- **Health conditions** : Blood pressure, Diabetes, etc.
- **Personal info** : Income, Family history, etc.

02
Preprocessing



Preprocessing

Data cleaning

Removing duplicated and invalid data

Train-test split

Data is splitted to train and test sets (66.7% / 33.3%).

Label encoding

Encoding nominal and ordinal features accordingly

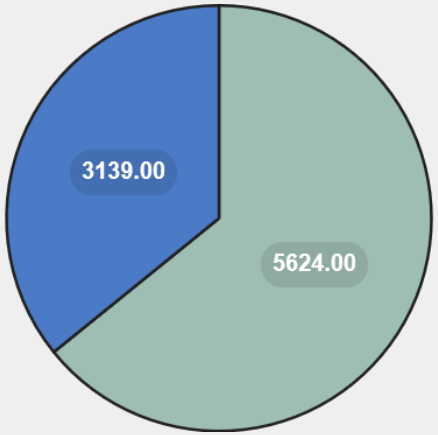
Normalizing

Scaling data for better ML convergence speed

03

Data Exploration

Distribution of Heart Attack Risk



Healthy
People with no risk of heart attack.

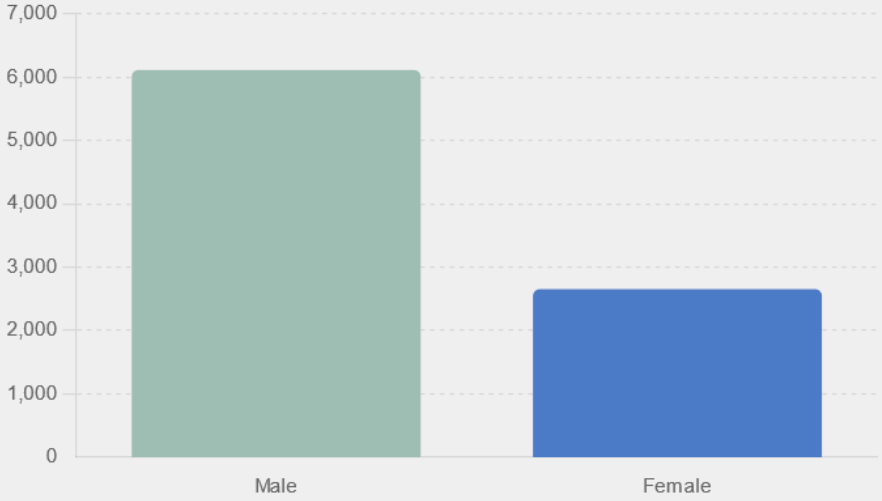
64.2%

Social withdrawal
People with heart risk of heart attack

35.8%

Distribution of the target label.

Distribution of Gender



Proportion of men and women in the dataset

Male 69.7%

Female 30.3%

Famous well-being centers

South Korea

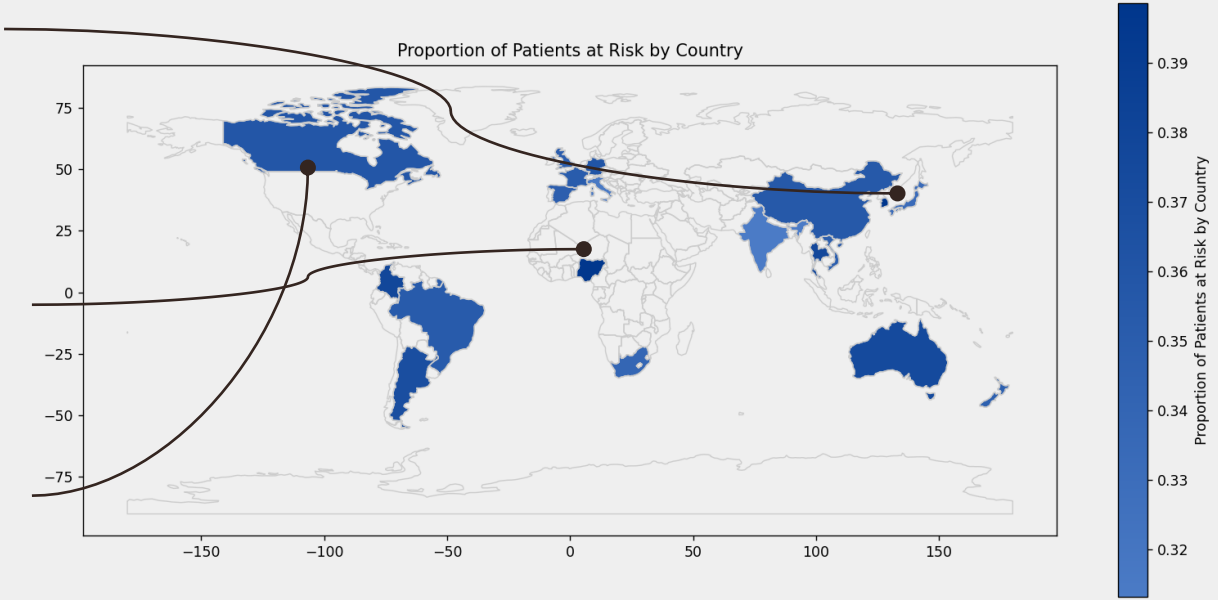
With 39.85% is the leading country

Nigeria

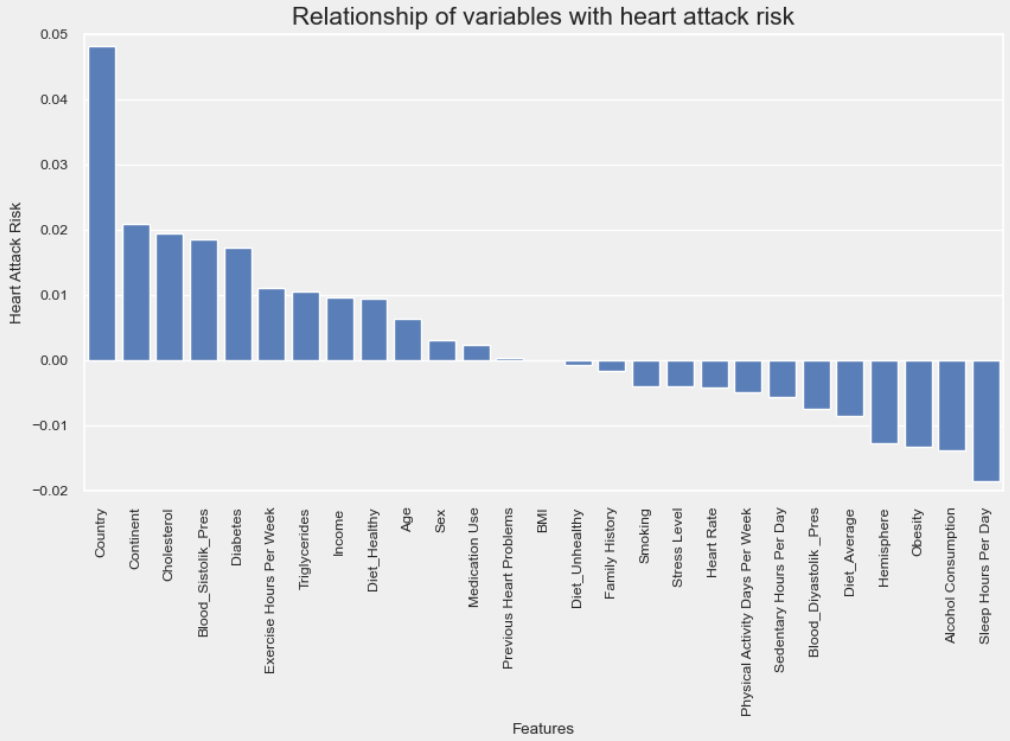
With 39.72% in second place

United States

With 39.52% in third place



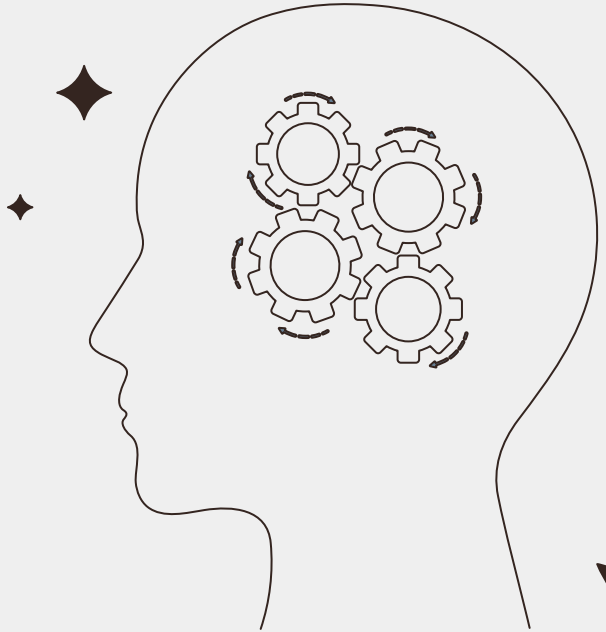
Analyzing Features Linked to Heart Attack Risk



04
Learning

Learning

- Different classification models are employed.
- Validation techniques are used to assess the model performance.



Classification Models

Logistic Regression

Descision Tree

Random Forest

SVM classifier

KNN

Guassian Naïve Bayes

First trial

Model	Accuracy	Recall
Logistic regression	64%	100%
SVM	64%	100%
Adaboost	64%	97%
KNN	56%	50%

To find a correct and reliable model with higher accuracy, it is vital to tune the hyperparameters.

Second trial using grid search

Model	Accuracy	Recall
Logistic regression	65%	100%
SVM	64%	100%
Adaboost	64%	99%
KNN	61%	87%

Duplicated Information can be detrimental !

Model Learning with Dimension Reduction

Preprocessing

Cleaning dataset as before

Dimension reduction

Performing dimension reduction using PCA method

Learning

Learning process of different classification models

Validation

Stratified K-Fold

Model Evaluation

Second trial using PCA

Model	Accuracy	Recall
Logistic regression	65%	100%
SVM	65%	100%
Adaboost	64%	99%
KNN	61%	88%

Validation

- Using Stratified K-Fold, the accuracy and performance are assessed.
- Validation results show that the accuracy for each model is not biased towards the training data.

100%

Percentage of patients that are predicted correctly.

65%

Accuracy up until now

Why is the accuracy limited ?

- Poor data quality
- Information is hidden → requires feature engineering
- Intrinsic difficulty of the problem → requires more data collection

05
Future Steps

Feature Engineering

Technical Knowledge

To ensure the discovery of useful features, it's crucial to gain technical medical knowledge about heart attacks or seek assistance from specialists.

EDA

More exploration is required to uncover relationships and patterns.

Resources

Dataset : Heart Attack Prediction

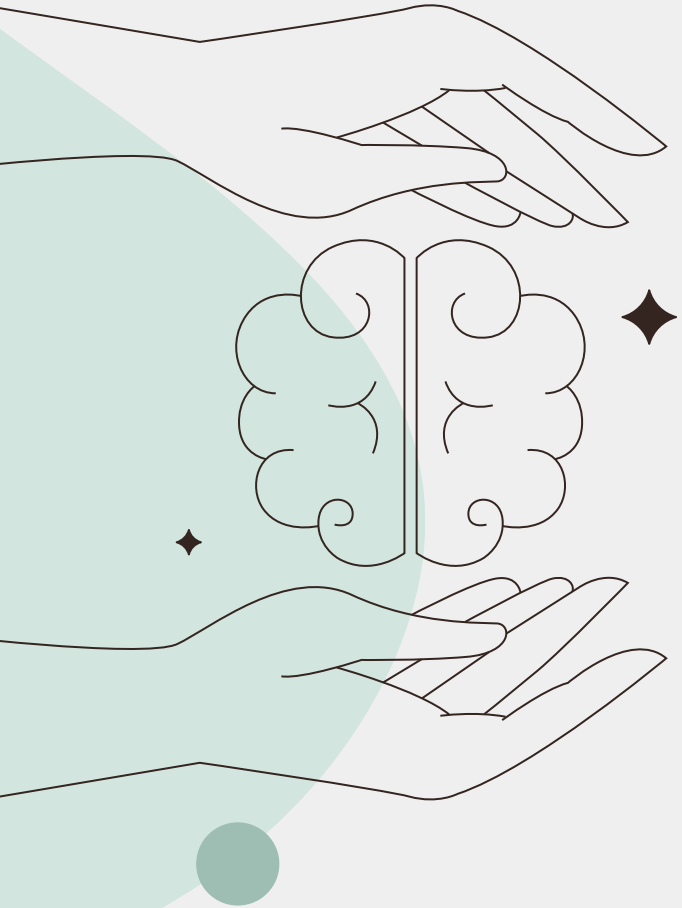
- https://www.kaggle.com/datasets/m1relly/heart-attack-prediction

Mahcine Learning methods

- Dr.Abolghasemi, Dr.Tavassolipoor, Machine Learning lecture notes

other resources:

- World Health Organization : Cardiovascular diseases
- Ourworldindata : Cardiovascular Diseases statistics



Thanks!

Do you have any questions?