Maya Hany Elshweiky 900204233

Cardiovascular Diseases Risk Prediction Dataset Description

Data Source

The Cardiovascular Diseases Risk Prediction Dataset was collected in 2023 and obtained from Kaggle.com. This data was sourced from the Behavioral Risk Factor Surveillance System (BRFSS) by the World Health Organization (WHO). BRFSS is the first national telephone survey system to collect state-level data on health risk behaviours, chronic diseases and preventive service utilization.

Dataset Overview

The dataset contains records related to personal lifestyle factors, encompassing a wide range of attributes that could influence cardiovascular health. Among the features included are demographic information such as age, gender (Sex), medical history (Diabetes), and self-reported health status (General Health). Additionally, the dataset comprises various other attributes related to lifestyle choices, habits, and health indicators. It serves as a collection of diverse data points aiming to capture various aspects of an individual's life and health that could contribute to the risk of developing cardiovascular diseases.

Dataset Details

- **Number of Variables: 19**
    - Categorical Variables: 7
    - Numerical Variables: 12
- **Number of Observations: 308,854**

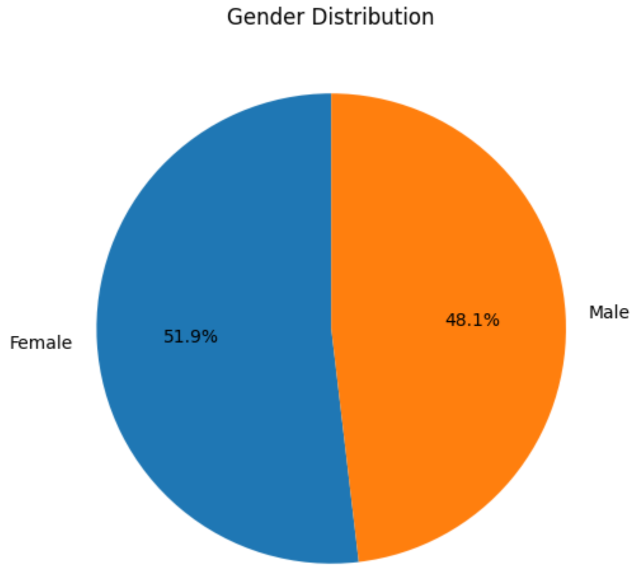
The variables and their descriptions are presented in the table below:

Variable	**Type**	**Description**	**Units of Measurement**
General_Health	Qualitative (Categorical)	General health condition of an individual	Categories ('Poor' 'Very Good' 'Good' 'Fair' 'Excellent')
Checkup	Qualitative (Categorical)	Time since last medical checkup	Intervals ('Within the past 2 years' 'Within the past year' '5 or more years ago' 'Within the past 5 years' 'Never')
Exercise	Qualitative (Categorical)	For past month, other than regular job, participation in any physical activities or exercises	Binary ('Yes' or 'No')
Heart_Disease (target variable)	Qualitative (Categorical)	Indicates presence of heart disease	Binary ('Yes' or 'No')
Skin_Cancer, Other_Cancer, Depression, Arthritis	Qualitative (Categorical)	Presence or absence of corresponding condition	Binary ('Yes' or 'No')
Diabetes	Qualitative (Categorical)	Respondents that reported having a diabetes. If yes, what type of diabetes it is/was.	Categories ('Yes' 'No' 'Yes, but female told only during pregnancy' or 'No, pre-diabetes or borderline diabetes')
Sex	Qualitative (Categorical)	Gender of the individual	Categories ('Male' or 'Female')
Age_Category	Qualitative (Categorical)	Categorization of individuals into age groups	Categories ('70-74' '60-64' '75-79' '80+' '65-69'

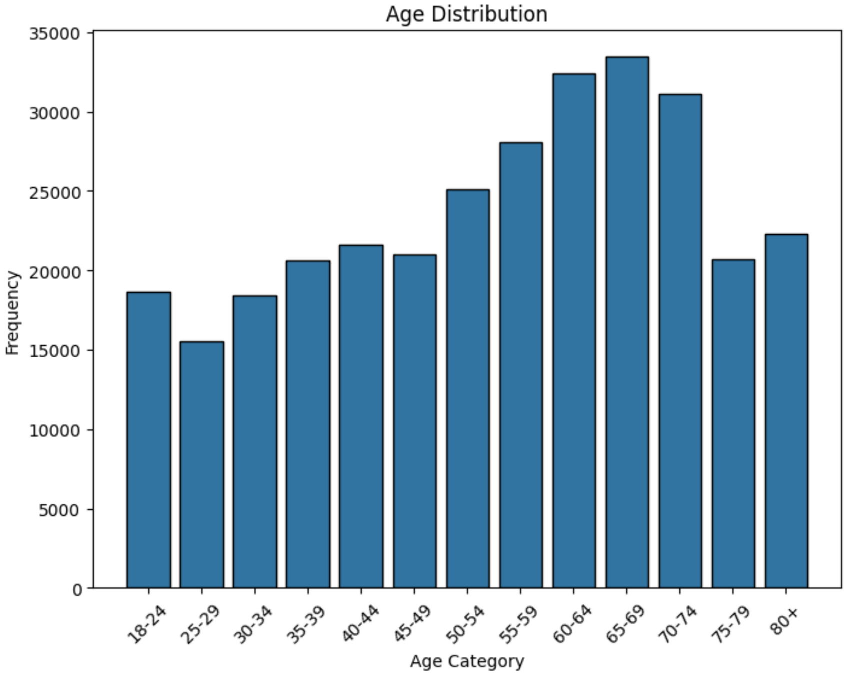
			'50-54' '45-49' '18-24' '30-34' '55-59' '35-39' '40-44' '25-29')
Height_(cm)	Quantitative (Numeric)	Height of the individual in centimetres	Centimetres
Weight_(kg)	Quantitative (Numeric)	Weight of the individual in kilograms	Kilograms
BMI	Quantitative (Numeric)	Body Mass Index gives an indication of whether an individual's weight is healthy for their height	Derived value from weight and height
Smoking_History	Qualitative (Categorical)	Description of smoking habits of the individual	Binary ('Yes' or 'No')
Alcohol_Consumption	Quantitative (Numeric)	Average units of alcohol consumed per week	Units per week
Fruit_Consumption, Green_Vegetables_Consumption, FriedPotato_Consumption	Quantitative (Numeric)	Number of consumption in a month	Units per month Example: a 128 response would be a representative of a person who consumes vegetables at least 4-5 times a day

Data Analysis and Visualization

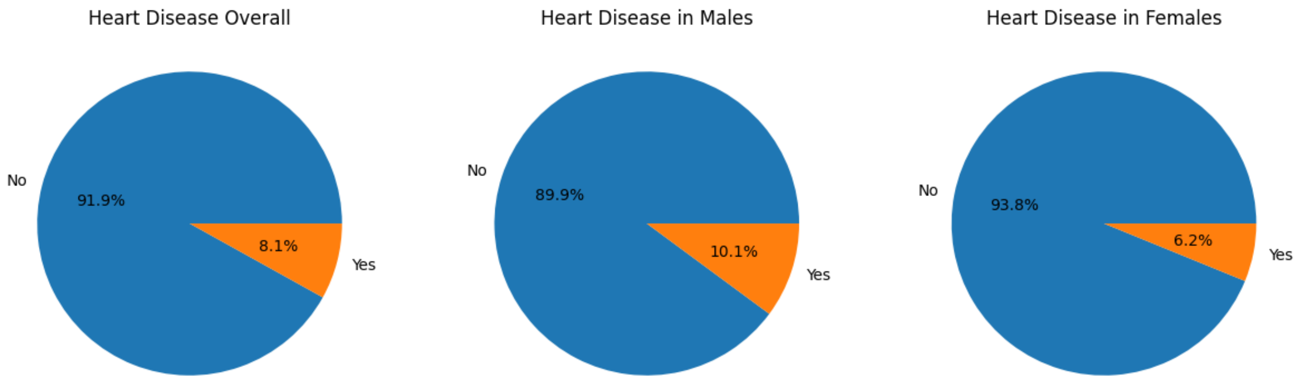
Analysis of Different Features



For gender, there is a balance between male and female, in which 51.9% are female and 48.1 are male.



As shown in the age distribution histogram, a significant proportion of the records are from ages 50 to 74. To illustrate this, between the age categories of 18 to 34, there are around 55,000 registrations. Between the age categories of 60 and 74, there are approximately 95,000 records.



As shown in the pie charts above, the first pie chart represents heart disease, which is the target variable; 25,017 patients have heart disease, which represents 8.1% of the data, and 283,837 patients who do not have heart disease, which represents 91.9% of the data. This shows that there is a class imbalance in the target variable. As shown in the second and third pie charts, the heart disease percentage of patients by gender, we can also see that men, which represents 10.1%, are more affected than women, which represents 6.2%.

Source and Citation

Source: Cardiovascular Diseases Risk Prediction Dataset

Citation: Alphiree. (2023). *Cardiovascular diseases risk prediction dataset*. Retrieved from https://www.kaggle.com/datasets/alphiree/cardiovascular-diseases-risk-prediction-dataset/data