Cardiovascular Diseases Risk Prediction

Project by Roberta Solom and Kadri-Ketter Kont

Dataset: Cardiovascular Diseases Risk Prediction

Repository: Github

Business understanding

Cardiovascular diseases (CVDs), responsible for 17.9 million annual deaths globally, encompass conditions like coronary heart disease and strokes. Unhealthy diet, physical inactivity, tobacco use, and excessive alcohol consumption are key behavioral risk factors. These factors contribute to raised blood pressure, glucose, lipids, and obesity, indicating elevated risks. Identifying high-risk individuals and ensuring access to appropriate treatments in primary health care facilities can prevent premature deaths and promote overall cardiovascular health.

Business goals. Our goal is to explore how people's lifestyle choices impact the risk of Cardiovascular Disease (CVD) and develop predictions based on these factors. We aim to use this information for, providing individuals with insights to proactively manage and reduce their CVD risk.

Business success criteria. Our business success depends on creating a really accurate model for predicting Cardiovascular Disease risk. We want to make sure it rarely misses identifying people at risk (lowest false negative rate), ensuring we catch potential issues early and make the predictions as reliable as possible.

Inventory of resources. Our resource inventory includes a comprehensive dataset featuring columns such as General Health, Checkup, Exercise, Heart Disease, Skin Cancer, Other Cancer, Depression, Diabetes, Arthritis, Sex, Age Category, Height (cm), Weight (kg), BMI, Smoking History, Alcohol Consumption, Fruit Consumption, Green Vegetables Consumption, and Fried Potato Consumption. Plus, we're equipped with Python, Jupyter Notebook, and essential Python libraries to analyze the data and build models effectively.

Requirements, assumptions and constraints. We'll stick to the requirements outlined by our course and ensure we follow them carefully throughout the project. We are going to use an open dataset from Kaggle.

Risks and contingencies. We've thought about things that could slow us down, and we have plans to deal with them. If we're short on time because of other important courses, we'll manage our tasks wisely and may ask for help. Given that the data is not collected by us, there is a possibility of containing inaccuracies. We will exercise caution to avoid misinterpreting any information during our analysis.

Terminology. We've put together a list of important words for our project, especially focusing on explaining what each column in our datasets means.

Columns	Definition
General_Health	Overall health assessment question.
Checkup	Inquiry about the last routine doctor visit duration.
Exercise	Non-work physical activities in the past month..
Heart_Disease	Respondents that reported having coronary heart disease or mycardialinfarction.
Skin_Cancer	Respondents that reported having skin cancer.
Other_Cancer	Respondents that reported having any other types of cancer.
Depression	Respondents that reported having a depressive disorder (including depression, major depression, dysthymia, or minor depression).
Diabetes	Respondents that reported having diabetes.
Arthritis	Respondents that reported having Arthritis.

Sex	Respondents gender.
Age_Category	Which age category respondent fits in.
Height_(cm)	Respondents height in centimeters.
Weight_(kg)	Respondents weight in kilograms.
BMI	Respondents BMI
Smoking_History	Inquires about smoking habits.
Alcohol_Consumption	About respondents alcohol consumption habits.
Fruit_Consumption, Green_Vegetables_Consumption, FriedPotato_Consumption	Capture dietary habits related to fruit, green vegetables, and fried potatoes.

Costs and benefits. For our personal development and school project in data mining, we're not dealing with money costs and gains. Our primary investment is time, but consider it as time dedicated to gaining interesting knowledge.

Data-mining goals. Our primary objective is to create various visual representations, such as plots, to illustrate the impact of different lifestyle factors on the presence of cardiovascular diseases. Furthermore, we're working on building an accurate prediction model and this model will be key in predicting cardiovascular disease risks.

Data-mining success criteria. Our data mining success depends on achieving high recall to accurately identify and diagnose a substantial number of cases, reducing the risk of false negatives. Prioritizing accuracy is equally crucial, ensuring correct identification of both positive and negative instances. This balanced approach, including correctly identifying cases without the condition, contributes to a robust and reliable model.

Data understanding

For this project we have outlined following **data requirements**:

Columns	Time range	Data format
General_Health	At the moment	Very Good Good Excellent Fair Poor
Checkup	Recent years	Within the past year Within the past 2 years Within the past 5 years 5 or more years ago Never
Exercise	Past month	True or false
Heart_Disease	Recent years	True or false
Skin_Cancer	Recent years	True or false
Other_Cancer	Recent years	True or false
Depression	Recent years	True or false
Diabetes	Recent years.	No Yes No, pre-diabetes or borderline diabetes Yes, but female told only during pregnancy
Arthritis	Recent years	True or false
Sex	At the moment	Female or male
Age_Category	At the moment	65-69 60-64 70-74

		55-59 50-54 80+ 40-44 45-49 75-79 35-39 18-24 30-34 25-29
Height_(cm)	At the moment	Centimeters
Weight_(kg)	At the moment	Kilograms
BMI	At the moment	BMI index
Smoking_History	Recent years	True or false
Alcohol_Consumption	Past month	Number of consumption in a month
Fruit_Consumption, Green_Vegetables_Consumption, FriedPotato_Consumption	Past month	Number of consumption in a month

Data availability. We have secured the necessary data from an open dataset obtained from Kaggle, which originates from the Centers for Disease Control and Prevention. The dataset is derived from the Behavioral Risk Factor Surveillance System (BRFSS), recognized as the nation's leading system for health-related telephone surveys. BRFSS systematically collects state data on U.S. residents, providing comprehensive insights into their health-related risk behaviors, chronic health conditions, and utilization of preventive services. This established and reputable source assures the availability and reliability of the data for our analysis.

Selection criteria. We're using a dataset from Kaggle, a clean CSV file about Cardiovascular Disease (CVD) that's 32.45 MB. It's our only data source, with 19 columns, and each column is vital for understanding CVD risks. All these columns help us find the answers we need for our project goals.

Describing data. Our CVD dataset from Kaggle is a well-organized and cleaned 32.45 MB CSV file with a substantial 309k rows, ensuring plenty of data for analysis. With 19 columns, including General_Health, Checkup, Exercise, and more, each field contributes uniquely to understanding CVD risks. The dataset aligns perfectly with our data-mining goals, covering essential variables for our analysis. All the expected fields are present, making the dataset suitable for our project. Its size and the richness of information in each column make it ideal for in-depth analysis and modeling. In summary, it's a solid foundation for exploring cardiovascular health and risks.

Exploring data. Since the data has been cleaned, it's more reliable. The only columns we see with data quality concerns are Fruit_Consumption, Green_Vegetables_Consumption, FriedPotato_Consumption, and Alcohol_Consumption. This is because the data represents the number of times something is consumed in a month. For instance, it could indicate how many times someone ate an apple. However, the challenge is that people eat different amounts, making it difficult to compare.

Data quality. The data quality is solid because the dataset is clean and fits well with what we want to achieve in data mining. It aligns well with our goals of studying cardiovascular health and risks. In simpler terms, the data seems trustworthy for our analysis.

Project plan

0. Starting the project

Tasks: Collecting information, finding an appropriate dataset, slides for the presentation, this document.
Hours: Team Member 1 (6 hours), Team Member 2 (6 hours).
Methods and Tools: Google, Kaggle, course materials (homeworks, practicals etc)

1. Data Exploration and Cleaning:

Tasks: Examine and clean the dataset, exploring relationships between variables.
Hours: Team Member 1 (5 hours), Team Member 2 (5 hours).
Methods and Tools: Python (Pandas, NumPy), data visualization libraries (Matplotlib, Seaborn).

2. Feature Engineering:

Tasks: Identify and transform relevant features for the prediction model.

Hours: Team Member 1 (8 hours), Team Member 2 (8 hours).

Methods and Tools: Python (Scikit-learn), statistical analysis.

3. Prediction Model Development:

Tasks: Select, train, and fine-tune machine learning algorithms.

Hours: Team Member 1 (10 hours), Team Member 2 (10 hours).

Methods and Tools: Machine learning algorithms (Random Forest, Logistic Regression), Scikit-learn.

4. Results Analysis, Visualization and Plot Making:

Tasks: Analyze model predictions, visualize results, and create informative plots.

Hours: Team Member 1 (10 hours), Team Member 2 (10 hours).

Methods and Tools: Python (Matplotlib, Seaborn), interpretability tools for machine learning models.

5. Report and Presentation:

Tasks: Make a project report and prepare a visually engaging presentation.

Hours: Team Member 1 (6 hours), Team Member 2 (6 hours).

Methods and Tools: Jupyter Notebook for report documentation, presentation tools (e.g., PowerPoint).

References

https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1

https://courses.cs.ut.ee/2023/ids/fall/Main/HomePage?action=download&upname=crisp_dm.pdf