Instructions for the project

Project: Using Exploratory Data Analysis and Visualization to predict Heart Diseases.

Client: OA-MEDCARE Think Tank

Terms of Reference

Introduction

Coronary Heart Disease (CHD) is the leading cause of death worldwide.[1] Predicting heart diseases based on a patient's characteristics has been the subject of a considerable amount of research over the years.[2-4] Several methodologies have been discussed in the literature. Traditional statistical methods are the basis of most predictive methods, such as the Cox proportional hazard model,[5] logistic models,[6], and other methods. More recently, machine learning and artificial intelligence methods have been employed.[7] However, there is still a debate about the clinical value of such approaches.[8]

This exercise aims to use exploratory data analysis (EDA) and data visualization to give a non-formal classification of the outcome of interest (presence vs absence of CHD) using the features available and to recommend a formal way to tackle the classification method at hand.

Data

The package provided by the OA-MEDCARE think tank for the analysis contains the following:

1- The dataset is in a CSV format.
2- The meta-data file: this file provides an overview of the data. A dictionary explains the variables and helpful links for understanding the meaning of the variables.

Objectives

The main objectives for the analysis that we seek your team to fulfil are:

1- To provide an overview for the data statistically and visually:
   a. Data discrepancies, issues, and any irregularities
   b. Summary statistics for the dataset after cleaning
   c. A copy for the final cleaned dataset
2- To provide an exploratory analysis for the dataset that summarizes the dataset after cleaning
3- To conduct an explanatory analysis that provides more in-depth insight into potential factors influencing the outcome of interest.

Deliverables

Your party (the analysis group) is expected to provide specific deliverables before noon on the 15th of January 2022. The deliverables as per this TOR should include the following:

Table 1: Deliverables' specifications.

Deliverable	Specifications
Technical Report	A full technical report that does not exceed four pages. The technical report should include an abstract of 150 words and an introduction about the report that does not exceed 200 words—a detailed section explaining the methodology used should follow. The results section should include the analytical results described in the methodology section. The results should be concise and answer the issues raised in the objective section of this ToR. Long tables or large graphs should be included as an appendix (Max. 5 pages).
A copy of the cleaned data set	The cleaned dataset should be attached with the deliverables in CSV format.
Analytical report	The analytical report should summarize the results from the exploratory and explanatory analysis. This report will include 200 words maximum executive summary at the beginning. The second part of the report will provide a brief overview of the work under the introduction section. The third part will include a discussion that includes: 1- Main statistical features of the variables in the dataset.

	2- Central statistical relationships and correlations 3- Data visuals included in the analytical part (important figures or central ideas that need specific visualizations can be included as necessary) 4- Discussion of the explanatory analysis about the potential factors that influenced the outcome of the interest. Finally, the report will include as a final section your recommendations and what further work is needed to take the analysis to the next step. The analytical report should not exceed five pages.
Presentation	The final deliverable will be a presentation, and the presentation should not exceed 15 minutes. Afterwards, there will be a 10-minute Q&A session regarding the presentation and the analysis.

Deadline and other notes

1- There is no restriction for the statistical and visualization package that you will use for this analysis. Preferably:
   a. R language
   b. Power BI
   c. Python
2- The code and the files used in the analysis with complete steps also must be submitted.
3- The deadline is at noon on the 15th of January 2022.
4- The analysis will not be considered if external parties are involved (the group should entirely do the work).
5- The font used Arial size 12 for the body text, 14 bold for the section names, and 12 bold for the subsection names.
6- The reports should have a 1-inch margin from all sides.
7- You can use landscape style for graphs that need larger width.
8- You should use 1.5 spacing between the line in the body text and single spacing inside the tables.
9- All tables and figures should be appropriately captioned.
10- The table caption should be placed directly above the table.
11- The figure caption should be placed under the figure directly.

12- You can use up to 3 levels of subsections as follows:

1. Main Section Heading

1.1 subsection level

1.1.1 subsection level

13- The technical and analytical reports should have separate title pages, and you can include them in the same file if separated.

Referencing and Main Title pages

1- The title page should include the project name, group names, the university logo, your course name, and your tutor name.
2- The second page should include the name of each team member and the detailed contribution of that member.
3- Using any references or quotations should be acknowledged in the work
4- Use Vancouver reference style in your work as used in this document or Harvard referencing style.

Thank you and Good luck

References

1. World Health Organisation. The top 10 causes of death Geneva: WHO; 2020 [Available from: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death.
2. Booth-Kewley S, Friedman HS. Psychological predictors of heart disease: a quantitative review. Psychological bulletin. 1987;101(3):343.
3. Funk M, Naum JB, Milner KA, Chyun D. Presentation and symptom predictors of coronary heart disease in patients with and without diabetes. The American journal of emergency medicine. 2001;19(6):482-7.
4. Blankstein R, Budoff MJ, Shaw LJ, Goff DC, Polak JF, Lima J, et al. Predictors of coronary heart disease events among asymptomatic persons with low low-density lipoprotein cholesterol: MESA (Multi-Ethnic Study of Atherosclerosis). Journal of the American College of Cardiology. 2011;58(4):364-74.
5. Empana J, Tafflet M, Escolano S, Vergnaux A, Bineau S, Ruidavets J, et al. Predicting CHD risk in France: a pooled analysis of the DESIR, Three City, PRIME, and SU. VI. MAX studies. European Journal of Cardiovascular Prevention & Rehabilitation. 2011;18(2):175-85.
6. Gordon T, Castelli WP, Hjortland MC, Kannel WB, Dawber TR. Predicting coronary heart disease in middle-aged and older persons: the Framington study. Jama. 1977;238(6):497-9.
7. Nakanishi R, Slomka PJ, Rios R, Betancur J, Blaha MJ, Nasir K, et al. Machine learning adds to clinical and CAC assessments in predicting 10-year CHD and CVD deaths. Cardiovascular Imaging. 2021;14(3):615-25.
8. Nusinovici S, Tham YC, Yan MYC, Ting DSW, Li J, Sabanayagam C, et al. Logistic regression was as good as machine learning for predicting major chronic diseases. Journal of clinical epidemiology. 2020;122:56-69.