Analysis of Heart Disease Data

Project Report

presented by
Club der toten Dichten (Team 12)
Marie-Christin Häge, 1913888
Finn Hülsbuch, 1913864
Thilo Dieing, 1692328
Lasse Lemke, 1914420
Eric Jacquomé, 1903834
Timotheus Gumpp, 1913876

submitted to the
Data and Web Science Group
Prof. Dr. Heiko Paulheim
University of Mannheim

December 2022

Contents

Chapter 1

Application Area and Goals

Heart disease is currently still one of the highest causes of mortality on earth (Nahar et al., 2013; Kavitha and Kannan, 2016; Statistisches Bundesamt, 2020). Given the successful application of data mining in other sectors e.g. banking and finance or marketing (Keleş, 2017) possible applications in the medical industry are plentiful. Yet the healthcare sector is information rich but knowledge poor (Soni et al., 2011). According to Soni et al. (2011) medical data sets provide great potential for data mining to be used in clinical diagnosis.

This aim of this project was the application of data mining methods, more specifically classification methods, to predict whether or not a patient could suffer from a heart disease. The successful application could help doctors and medical staff with diagnosing patients by automatically analysing historical test result data of the patient and give a prediction when a higher potential of heart problems arise. By doing this analysis patients flagged for potential heart disease could possibly be prioritised. Due to the immense amount of stress and long working hours medical personal are facing, having an additional instance looking at the data could be beneficial. In the past such approaches have already been tested and proven to be a good diagnostic option (Usha Rani, 2011). Jabbar, Deekshatulu, and Chandra (2013) state that data mining techniques answer several important and critical questions related to healthcare and that they can improve the provision of quality services to patients.

This project report is based on the "Heart Disease Data Set" (Janosi et al., 1988) which, despite its age is still relevant given the fact that it consists of results of medical tests. In addition to that the validity is assumed because it is frequently used in contemporary research (see Nahar et al., 2013; Usha Rani, 2011; Aha and Kibler, 1988).

- Heart disease is one of the highest causes of mortality (Nahar et al., 2013; Kavitha and Kannan, 2016; Statistisches Bundesamt, 2020)
- DM is successfully applied in other sectors e.g. banking and finance or marketing (Keleş, 2017) but healthcare is still "information rich" but knowledge poor". (Soni et al., 2011)
- explain how classification can be used in the medical field (explanation classification (Usha Rani, 2011))
- Aiding doctors with diagnosing patients by giving a classification
  - "Data mining techniques have been widely used in diagnostic and health care applications because of their predictive power. Data mining algorithms can learn from past examples in clinical data and model the oftentimes non-linear relationships between the independent and dependent variables. The resulting model represents formalized knowledge, which can often provide a good diagnostic opinion." ([Usha Rani, 2011, p. 2]) (Usha Rani, 2011)
  - "Classification is a pervasive problem that encompasses many diverse applications. To improve medical decision making data mining techniques have been applied to variety of medical domains. Many health care organizations are facing a major challenge is the provision of quality services like diagnosing patients correctly and administering treatment at reasonable costs. Data mining techniques answer several important and critical questions related to health care." ([jabbar et al., 2013, p. 86]) (Jabbar, Deekshatulu, and Chandra, 2013)
- Data Mining helps to extract patterns in the process of knowledge discovery. DM provides new techniques which help the humans to analyze and understand large amounts of data for difficult and unsolved problems. (Usha Rani, 2011)
- Validity of our dataset: it is often used in the scientific field. (Nahar et al., 2013; Usha Rani, 2011; Aha and Kibler, 1988)
- A faster and more precise detection of a possible heart disease will enable a more immediate treatment and thus may save more lives.

1.1 Goals

- Giving doctors and medical staff a prediction of the medical status of the patient to increase awareness. medical history data contains huge amounts of test results and can be out of the scope of the examination. A automatic classification of a patients test values could increase the doctors attention to a more holistic overview.
- Medical history data consists of a large number of tests required to diagnose a particular disease (Gupta, Kumar, and Bhatnagar, 2007)
- Using a supervised learning classification algorithm to learn from historical, labled data to derive a model that can classify new data.

Chapter 2

Structure and Size of the Dataset

2.1 Structure of the Dataset

2.2 Size of the Dataset

Chapter 3

Preprocessing

Chapter 4

Datamining

Chapter 5

Results

Bibliography

Nahar, Jesmin et al. (2013). "Association Rule Mining to Detect Factors Which Contribute to Heart Disease in Males and Females". In: *Expert Systems with Applications* 40.4, pp. 1086–1093. ISSN: 09574174. DOI: 10.1016/j.eswa.2012.08.028. URL: https://linkinghub.elsevier.com/retrieve/pii/S095741741200989X (visited on 11/07/2022).

Kavitha, R. and E. Kannan (2016). "An Efficient Framework for Heart Disease Classification Using Feature Extraction and Feature Selection Technique in Data Mining". In: *2016 International Conference on Emerging Trends in Engineering, Technology and Science (ICETETS)*. 2016 International Conference on Emerging Trends in Engineering, Technology and Science (ICETETS). Pudukkottai, India: IEEE, pp. 1–5. ISBN: 978-1-4673-6725-7. DOI: 10.1109/ICETETS.2016.7603000. URL: http://ieeexplore.ieee.org/document/7603000/ (visited on 11/06/2022).

Statistisches Bundesamt (2020). *Gestorbene: Deutschland, Jahre, Todesursachen, Geschlecht*. URL: https://www-genesis.destatis.de/genesis/online?operation=previous&levelindex=0&step=0&titel=Tabellenaufbau&levelid=1665392693797&acceptscookies=false#abreadcrumb (visited on 10/10/2022).

Keleş, Mümine KAYA (2017). "An Overview: The Impact of Data Mining Applications on Various Sectors". In: p. 5.

Soni, Jyoti et al. (2011). "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction". In: *International Journal of Computer Applications* 17.8, pp. 43–48. ISSN: 09758887. DOI: 10.5120/2237-2860. URL: http://www.ijcaonline.org/volume17/number8/pxc3872860.pdf (visited on 11/07/2022).

Usha Rani (2011). "Analysis Of Heart Diseases Dataset Using Neural Network Approach". In: *International Journal of Data Mining & Knowledge Management Process* 1.5, pp. 1–8. ISSN: 2231007X. DOI: 10.5121/ijdkp.2011.1501. URL: http://www.aircconline.com/ijdkp/V1N5/0911ijdkp01.pdf (visited on 11/06/2022).

Jabbar, M. Akhil, B.L. Deekshatulu, and Priti Chandra (2013). "Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm". In: *Procedia Technology* 10, pp. 85–94. ISSN: 22120173. DOI: 10.1016/j.protcy.2013.12.340. URL: https://linkinghub.elsevier.com/retrieve/pii/S2212017313004945 (visited on 11/06/2022).

Janosi, Andras et al. (1988). *UCI Machine Learning Repository: Heart Disease Data Set*. URL: https://archive.ics.uci.edu/ml/datasets/Heart+Disease (visited on 10/10/2022).

Aha, D and Dennis Kibler (1988). "Instance-Based Prediction of Heart-Disease Presence with the Cleveland Database". In: *University of California* 3.1, pp. 3–2.

Gupta, Anamika, Naveen Kumar, and Vasudha Bhatnagar (2007). "Analysis of Medical Data Using Data Mining and Formal Concept Analysis". In: 1.11, p. 4.